IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Damon H
Dec 10, 2022

Outline

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

Executive Summary

- Summary of methodologies
  - Data gathering from public SpaceX API and Wiki pages
  - Explored data with visualization techniques including dashboard generation and folium maps
  - Classified successful landings
  - Standardize data for best parameters for machine learning models
- Summary of all results
  - Generated four machine learning models with an approximate accuracy of 83%
    1. Logistic Regression
    2. Support Vector
    3. Decision Tree
    4. K-Nearest Neighbors

Introduction

Project background and context

- The commercial space age is here, companies are making space travel affordable for everyone and perhaps the most successful is SpaceX.
- One reason for SpaceX's success is their rocket launches are relatively inexpensive. SpaceX advertises Falcon9 rocket launches at a cost of $62 million where others cost upwards of $165 million and much of the savings is because SpaceX can reuse the first stage with a successful landing.
- Therefore, if we can determine whether the first stage will land, we can determine the cost of a launch.

Problem

- Space Y would like to compete with SpaceX and our role is to determine the price of each launch.

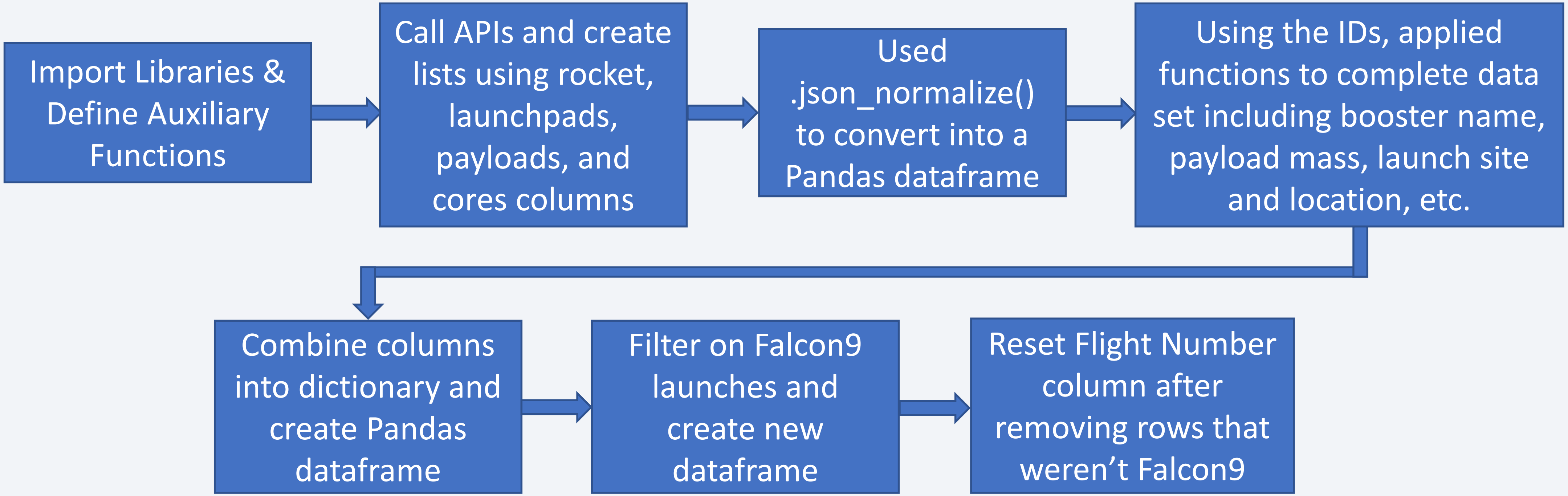
Section 1

Methodology

Methodology

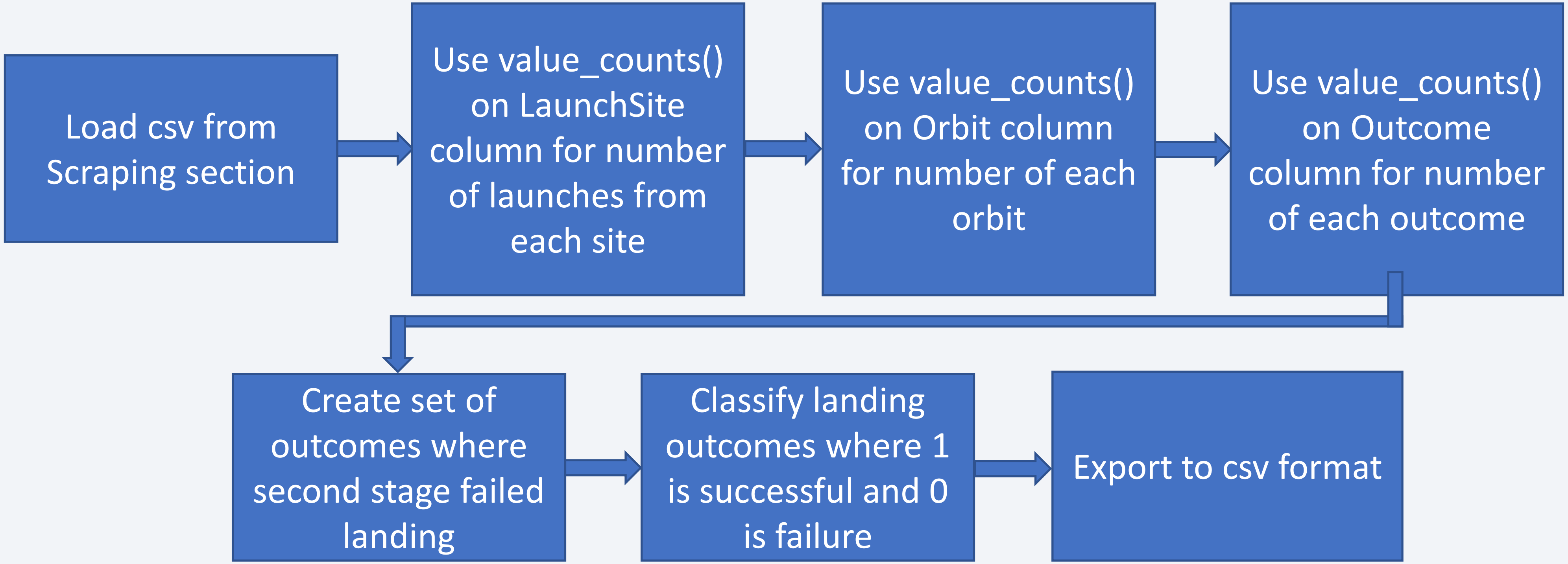
- Data collection methodology:
  - Data was gathered from the SpaceX REST API with data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications, and landing outcome.
- Perform data wrangling
  - Performed Exploratory Data Analysis (EDA) to find patterns and determine the labels for training supervised models.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Preprocessed to standardize the data and split it into training and testing data. Trained the model and performed Grid Search, yielding the hyperparameters for best performance. Determined the model with the best accuracy using the training data. Tested with Logistic Regression, Support Vector machines, Decision Tree Classifier, and K-nearest neighbors and output the confusion matrix

Data Collection – SpaceX API



[Github link to Coursera-Applied-Data-Science-Capstone Data Collection API Notebook]

Data Wrangling



[Github link to Coursera-Applied-Data-Science-Capstone Data Wrangling Notebook]

EDA with SQL

- Displayed the following…
  - Unique landing outcomes
  - Unique launch sites
  - 5 records where launch sites 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 of first successful landing outcome was achieved
  - Names of the boosters which have success in drone ships and payload masses greater than 4,000kg, but less than 6,000kg
  - Total number of successful and failed mission outcomes
  - Booster versions which have carried maximum payload mass
  - Names of the months, failed landings on drone ships, their booster versions, and lanuch sites for 2015
  - Count of successful landings between June 4, 2010, and March 20, 2017, in descending order

Github link to Coursera-Applied-Data-Science-Capstone EDA with SQL Notebook

EDA with Data Visualization

- Plots used include the following
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload Mass vs. Launch Site
  - Orbit vs. Success Rate
  - Flight Number vs. Orbit
  - Success Yearly Trend
- Scatter plots, line charts, and bar plots used to compare relationships between variables to determine whether a relationship existed tin order to enable training the machine learning model

[Github link to Coursera-Applied-Data-Science-Capstone EDA with Visualization Notebook]

Build an Interactive Map with Folium

- Folium maps marked Launch Sites, successful and unsuccessful landings, and proximity key locations: Railway, Highway, Coast, and City
- Folium maps help gain understanding on why launch sites may be located where they are and provides visualization of successful landings relative to location

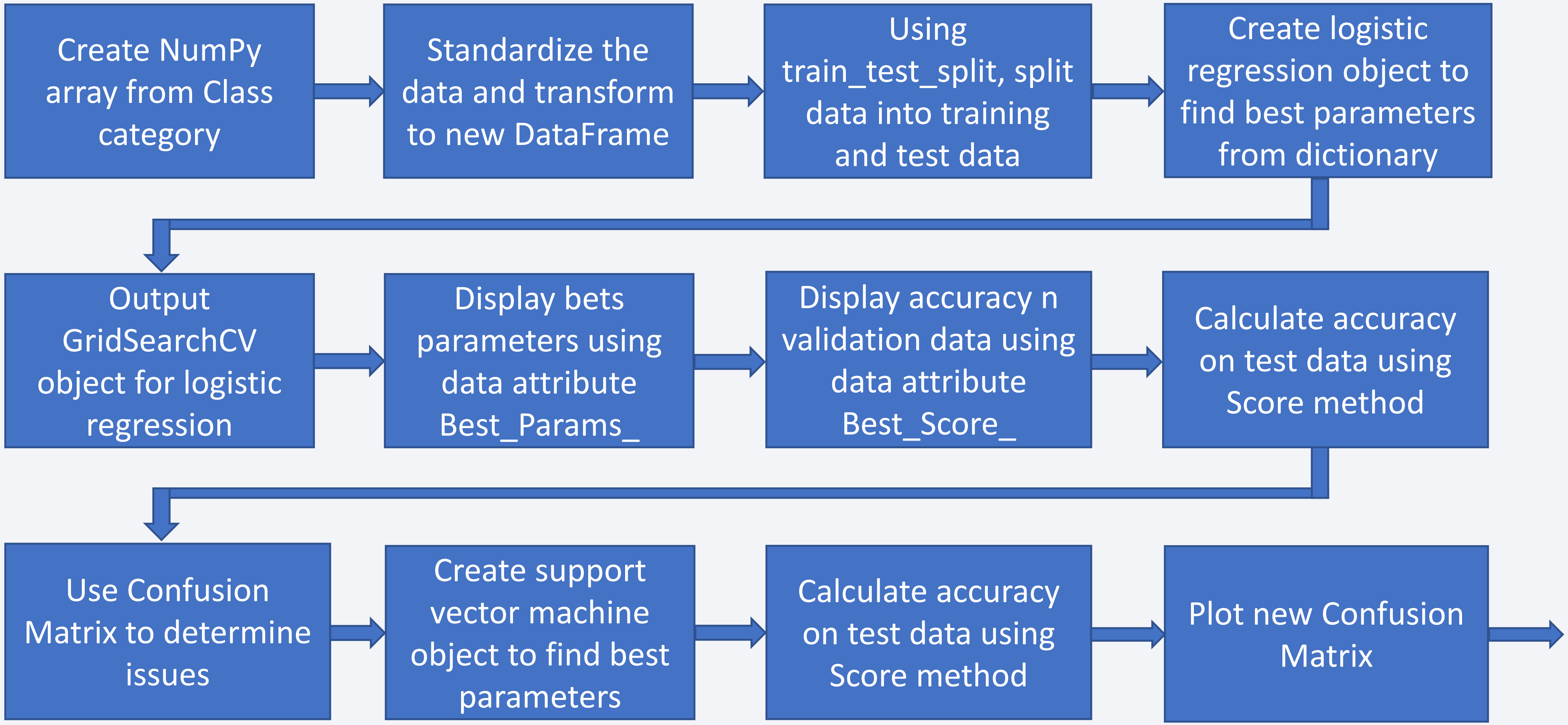
Github link to Coursera-Applied-Data-Science-Capstone Interactive Visual Analytics with Folium Notebook

Build a Dashboard with Plotly Dash

- Dashboard includes pie chart and scatter plot
  - Pie chart shows distribution of successful landings across launch sites and individual launch site success rates
  - Scatter plot shows variance of successes across launch sites, payload masses, and booster versions

Github link to Coursera-Applied-Data-Science-Capstone Dashboard Application with Plotly Dash Notebook

Predictive Analysis (Classification)



Predictive Analysis (Classification) – continued…

Create Decision Tree Classifier Object

Fit Object into best parameters from dictionary

Calculate accuracy of Tree on test data using Score method

Plot new Confusion Matrix

Create k Nearest Neighbor object

Fit Object into best parameters from dictionary

Calculate accuracy of Tree on test data using Score method

Plot new Confusion Matrix

Find best method from logistic regression, support vector machine, decision tree classifier, and k nearest neighbors

[Github link to Coursera-Applied-Data-Science-Capstone Machine Learning Predictions Lab Notebook]()

Results

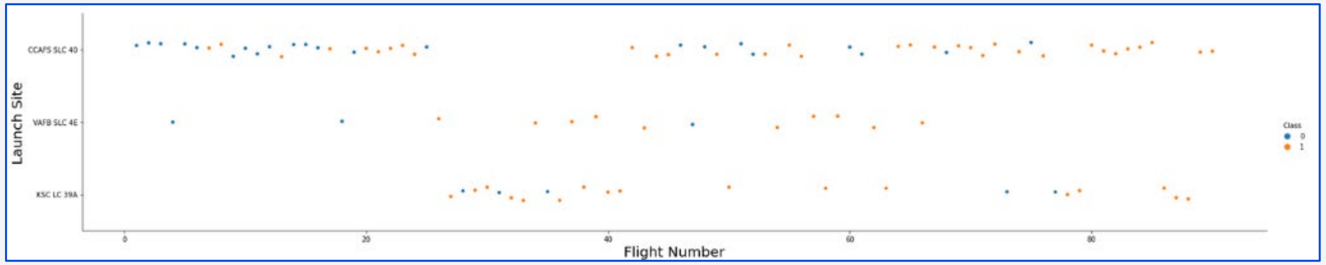
- Exploratory data analysis results with SQL
  - Total number of successful and failed mission outcomes

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

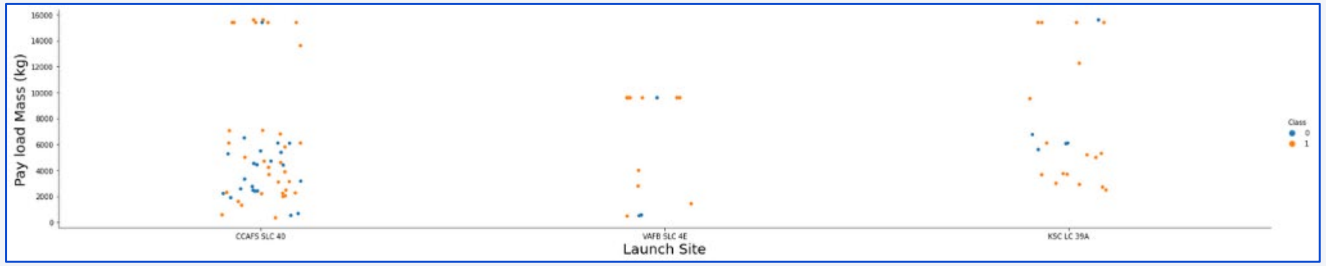
  - Landing outcomes between June 4, 2010, and March 20, 2017

landing__outcome	COUNT
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

- Exploratory data analysis visualization results
  - Flight Number vs. Launch Site

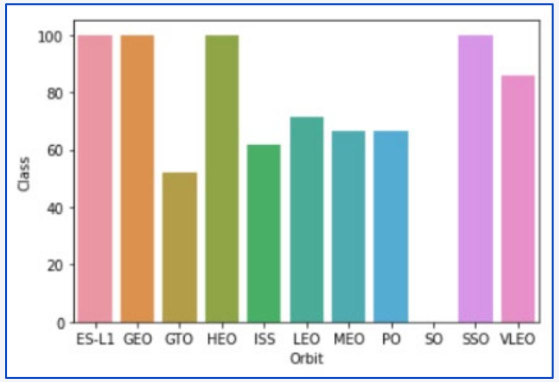


  - Payload vs. Launch Site

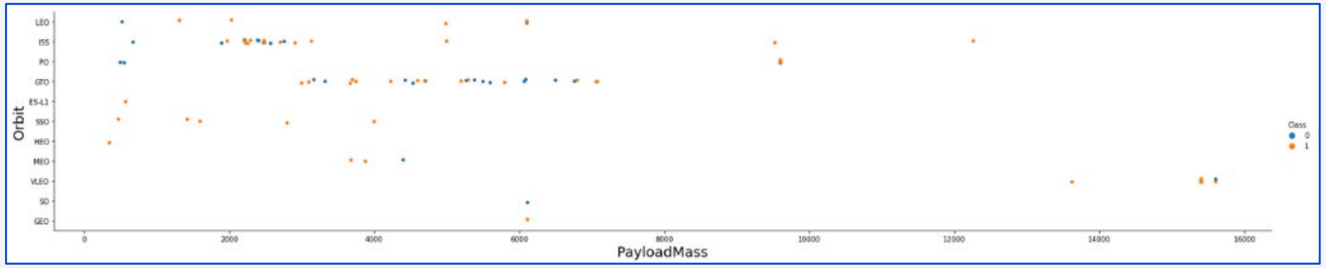


Results

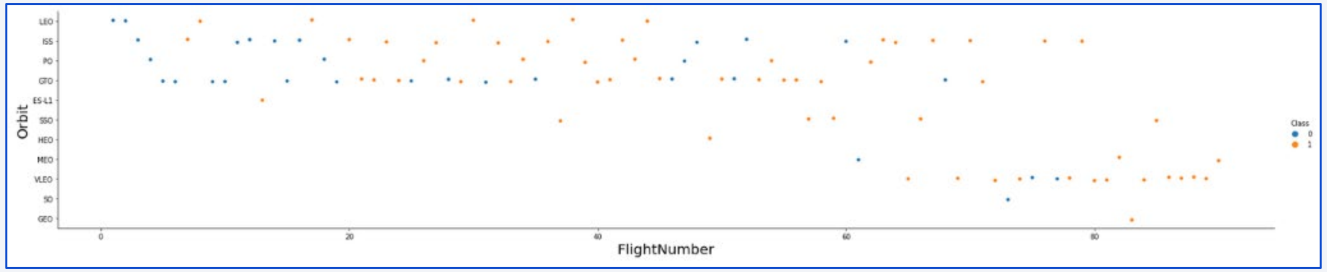
- Exploratory data analysis visualization results
  - Orbit by Class



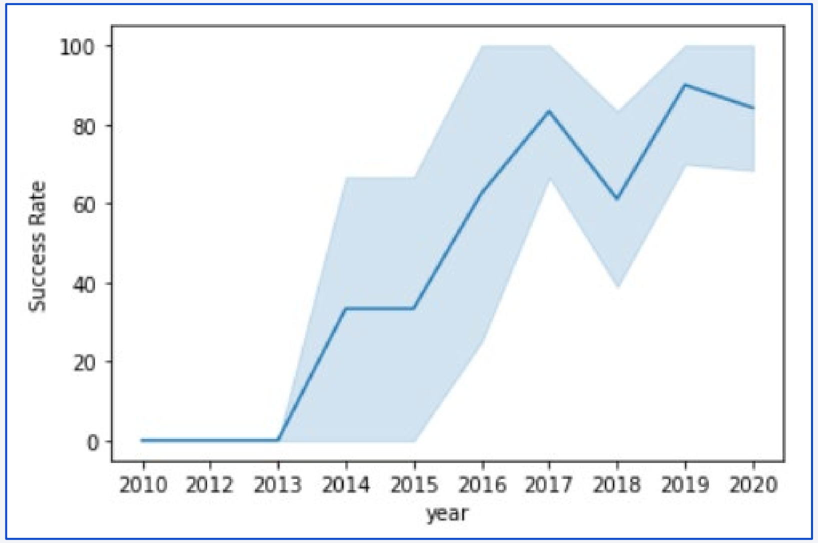
  - Payload Mass vs. Orbit



  - Flight Number vs. Orbit

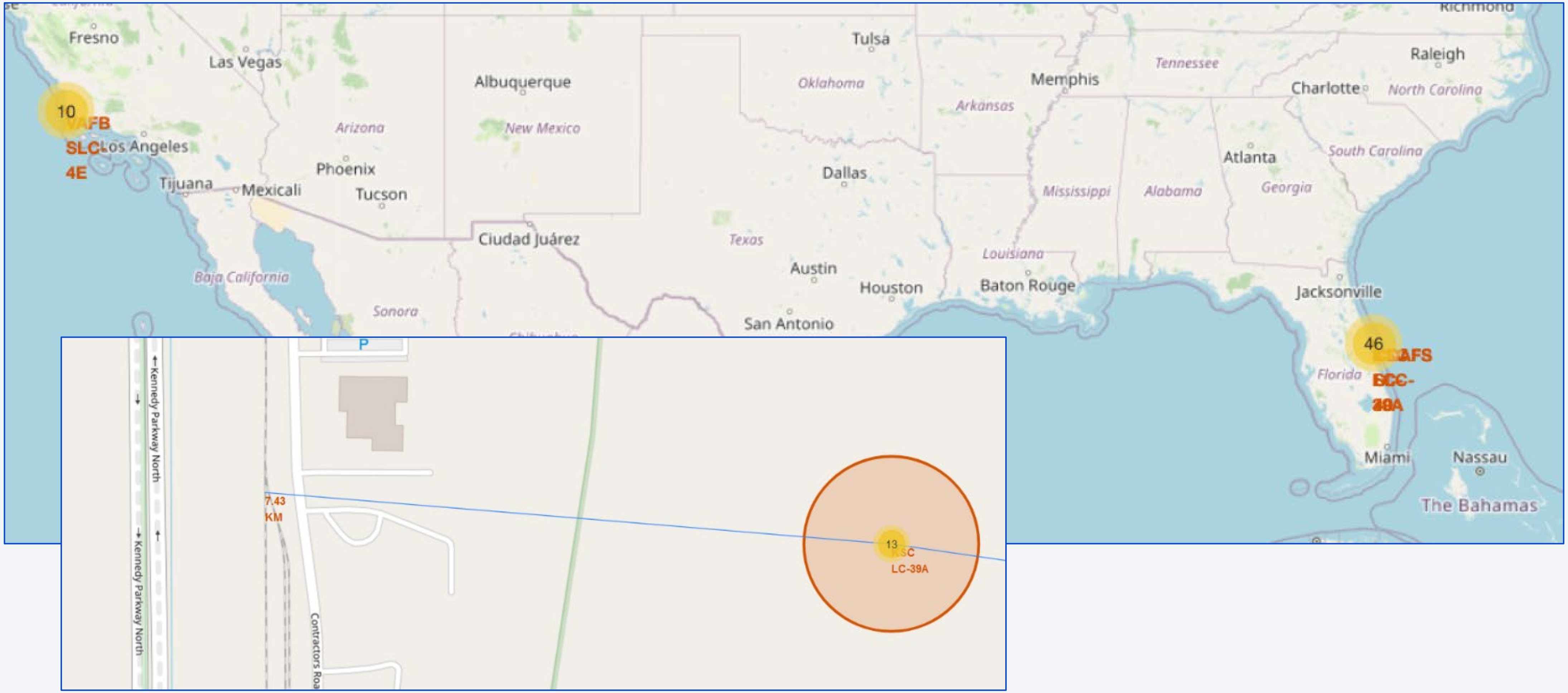


  - Yearly Trend



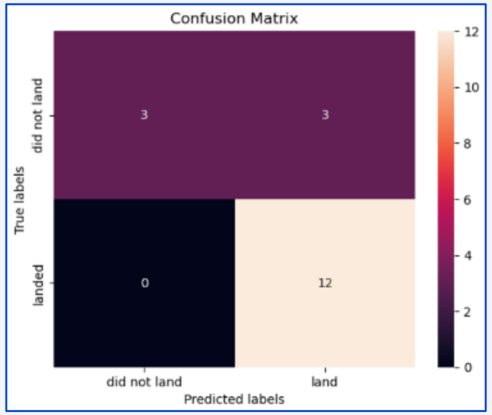
Results

- Interactive analytics demo in screenshots

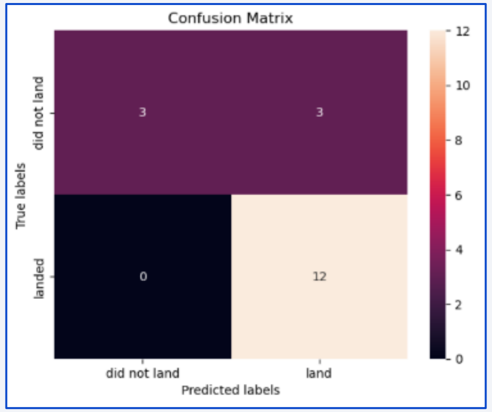


Results

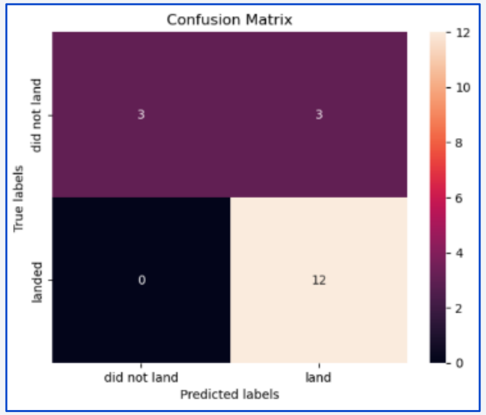
- Predictive Analysis Results
  - Accuracy of Logistical Regression: 83.3%



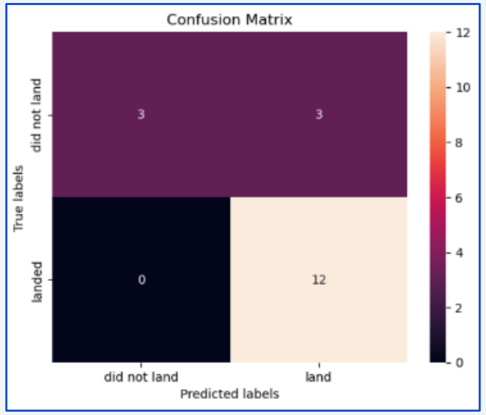
  - Accuracy of Support Vector Machine: 83.3%



- Best Method: Decision Tree Classifier
  - Accuracy of Decision Tree Classifier: 94.4%



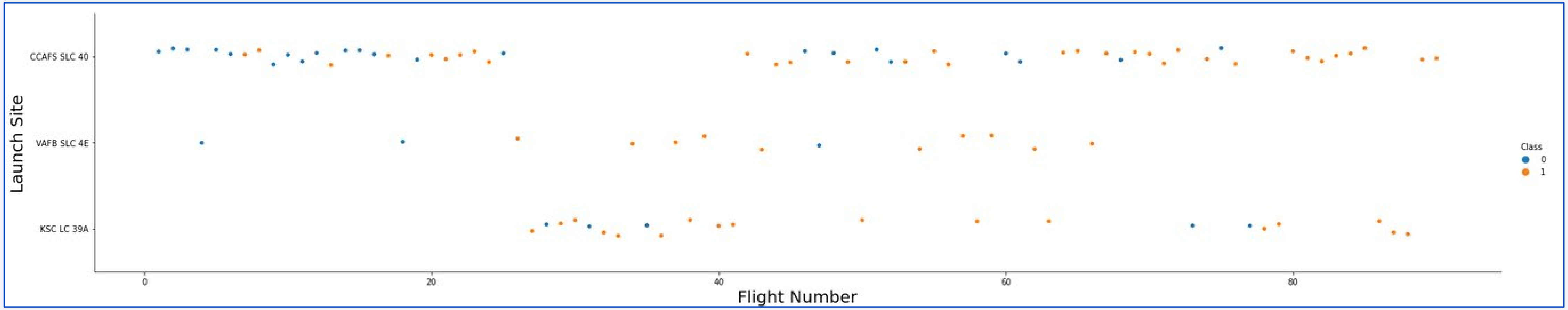
  - Accuracy of K-Nearest Neighbors: 83.3%



Section 2

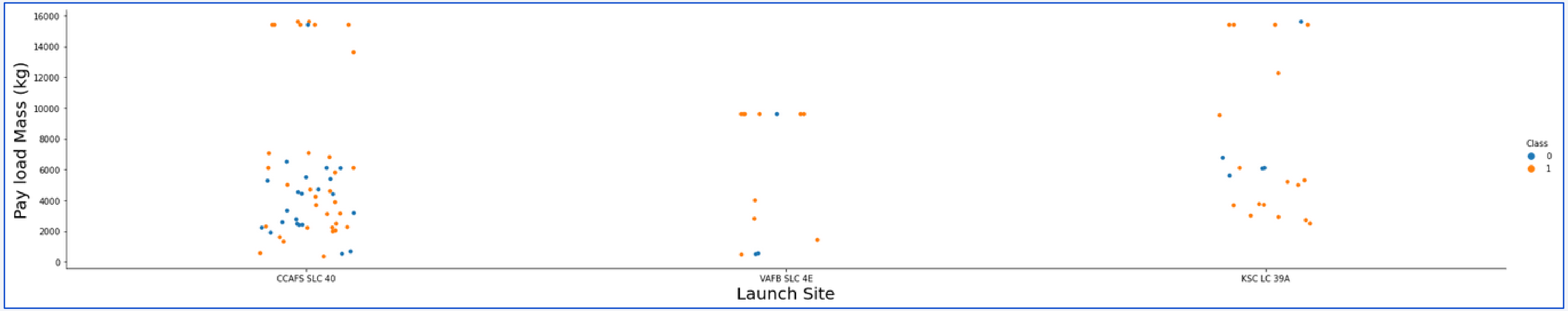
Insights drawn from EDA

Flight Number vs. Launch Site



- Most Launch Sites were CCAFS SLC-40
- There are streaks where a specific launch site was used often
  - Initially, it was CCAFS SLC-40, then KSC LC 39A, and then more at CCAFS SLC-40
- There appears to be more failed missions in the earlier flights than later as well

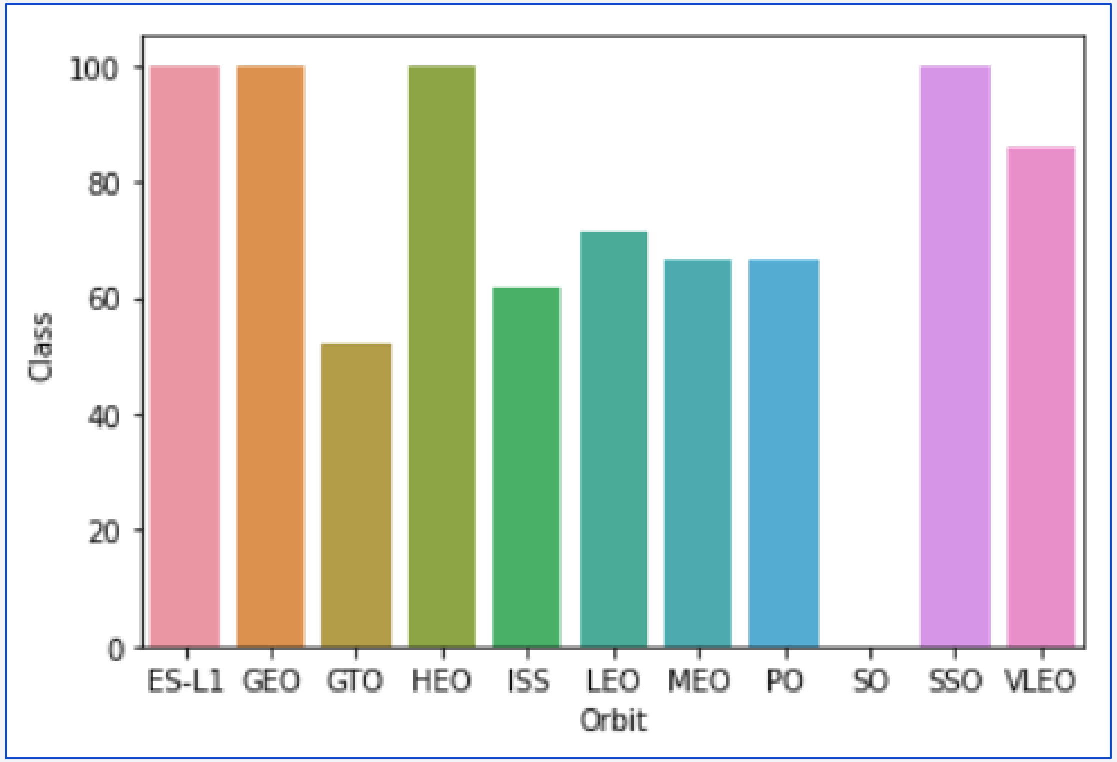
Payload vs. Launch Site



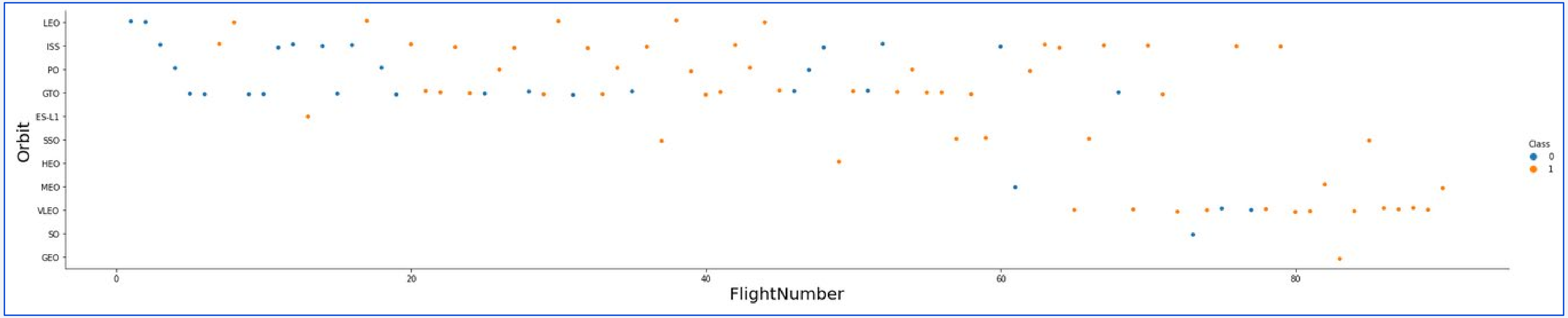
- There are no launches from VAFB SLC-4E with payloads larger than 10,000kg
- VAFB SLC-4E has the bulk of the launches with payloads of 10,000kg
- Higher success rates with the higher payloads as well
- CCAFS SLC-40 has the most launches

Success Rate vs. Orbit Type

- ES-L1, GEO, HEO, and SSO have the highest success rates
- GTO has the lowest success rate
- No launches to orbit SO

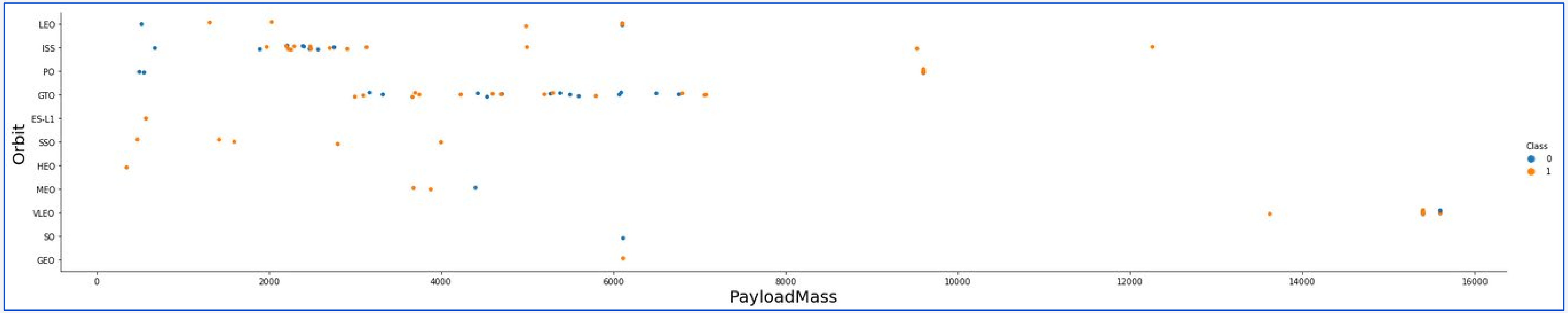


Flight Number vs. Orbit Type



- LEO Orbit had success after 2 flights
- GTO also increased in success overtime, but to a much less degree

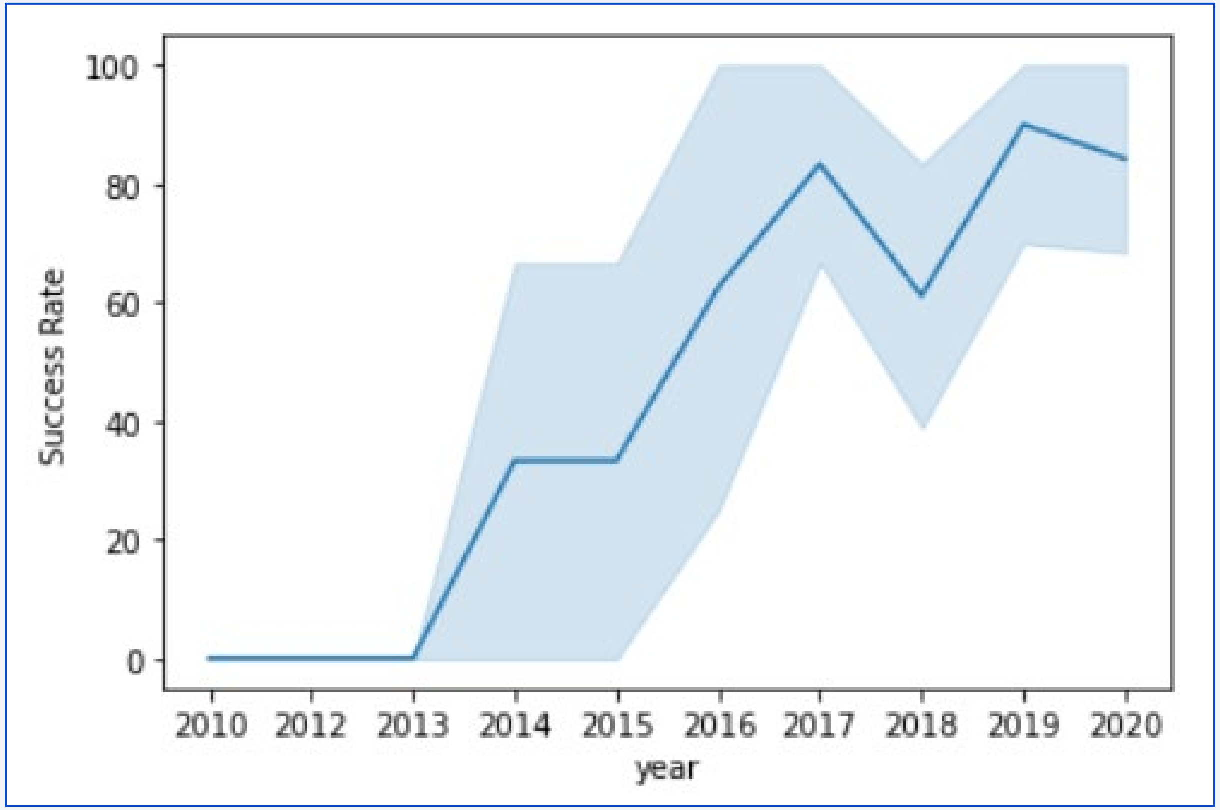
Payload vs. Orbit Type



- ISS had total success with higher payloads
- GTO is undistinguishable for success and fails relative to payload mass
- SSO only had success, but none with larger than 4,000kg payloads

Launch Success Yearly Trend

- Success rate dramatically increased from 2013 to 2017
- From 2017 to 2020, however, success rate only increased slightly



All Launch Site Names

- Unique Launch Sites

Launch Site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- Five records where launch sites begin with `CCA`

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

- Total payload carried by boosters from NASA
  - 45,596kg
  - %sql select sum(payload_mass__kg_) as sum from SPACEX where customer like 'NASA (CRS)'

Average Payload Mass by F9 v1.1

- Average payload mass carried by booster version F9 v1.1
  - 2,534kg

  - %sql select avg(payload_mass_kg_) as Average from SPACEX where booster_version like 'F9 v1.1%'

First Successful Ground Landing Date

- Date of the first successful landing outcome on ground pad
  - June 4, 2010
  - %sql select min(date) as Date from SPACEX where mission_outcome like 'Success'

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

  - %sql select booster_version from SPACEX where (mission_outcome like 'Success') AND (payload_mass_kg_ BETWEEN 4000 AND 6000) AND (landing_outcome like 'Success (drone ship)')

Total Number of Successful and Failure Mission Outcomes

- Total number of successful and failure mission outcomes
  - %sql SELECT mission_outcome, count(*) as Count FROM SPACEX GROUP by mission_outcome ORDER BY mission_outcome

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

Boosters Carried Maximum Payload

- Names of the booster versions which have carried the maximum payload mass
  - %sql select booster_version from SPACEX where payload_mass__kg_=(select max(payload_mass__kg_) from SPACEX)

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

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

MONTH	landing__outcome	booster_version	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

  - %sql select MONTHNAME(DATE) as Month, landing_outcome, booster_version, launch_site from SPACEX where DATE like '2015%' AND landing_outcome like 'Failure (drone ship)'

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

- Landing outcomes between June 4, 2010, and March 20, 2017, ranked in descending order

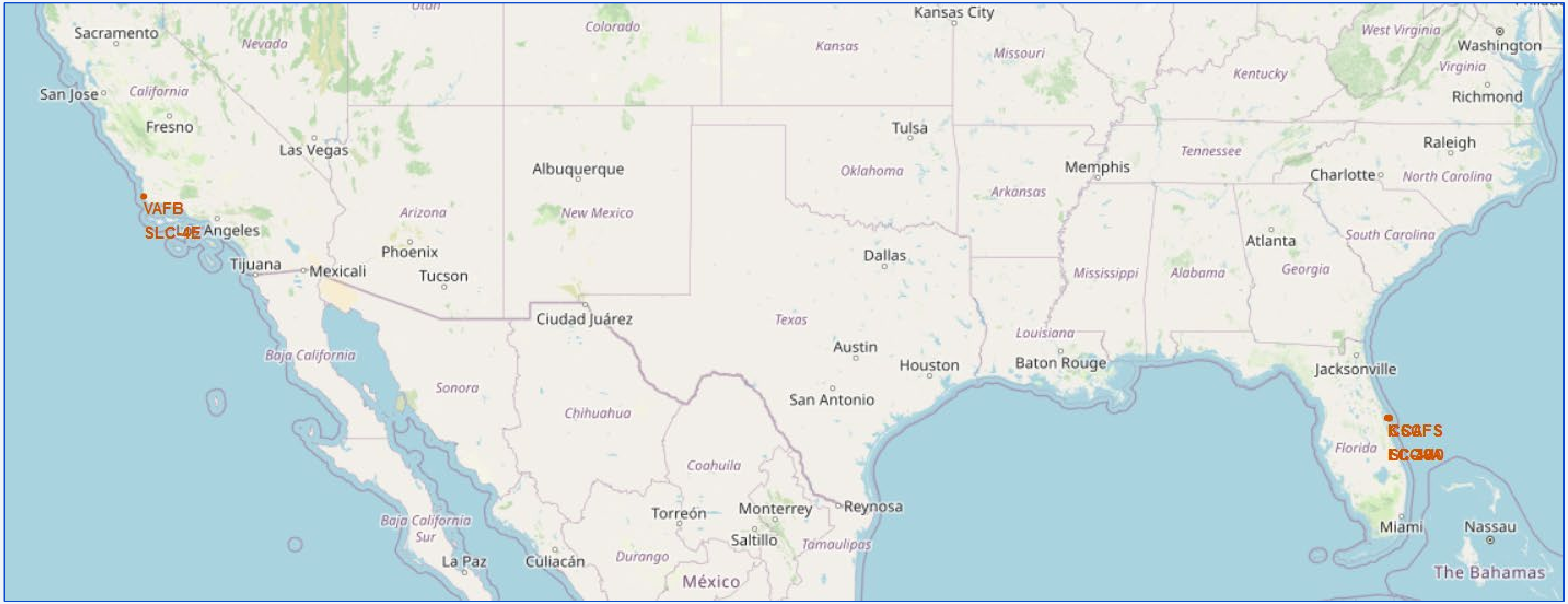
    - %sql select landing_outcome, count(*) as count from SPACEX where Date >= '2010-06-04' AND Date <= '2017-03-20' GROUP by landing_outcome ORDER BY count Desc

landing_outcome	COUNT
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

Section 3

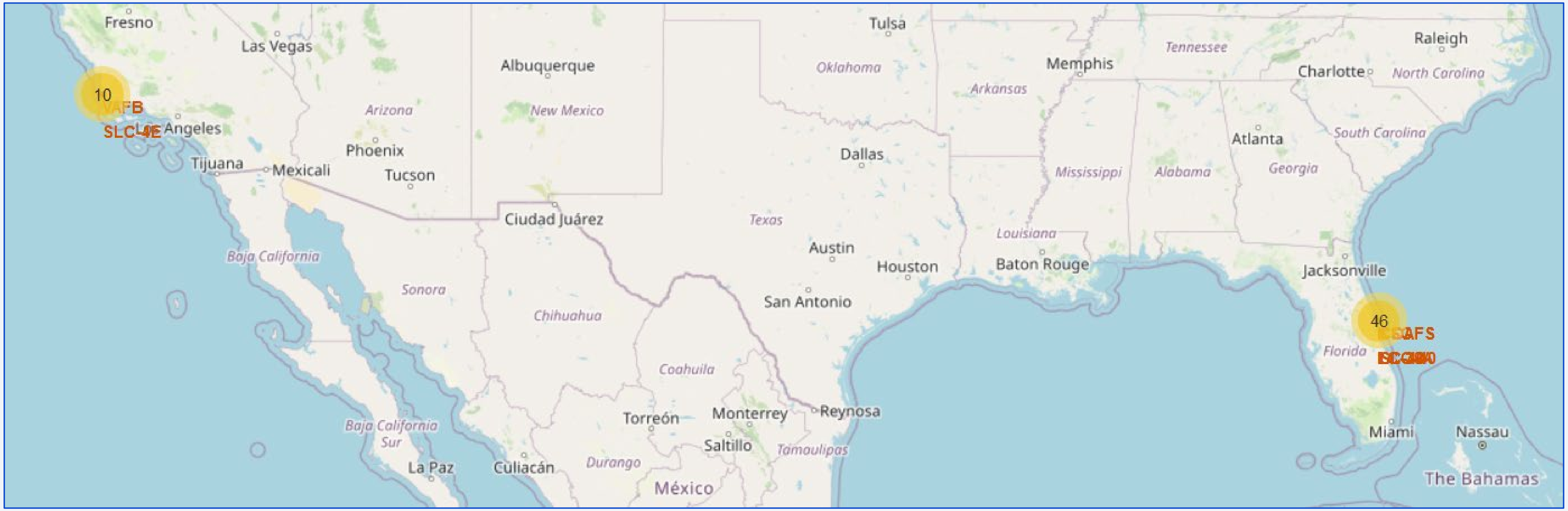
Launch Sites Proximities Analysis

Launch Site Location Markers



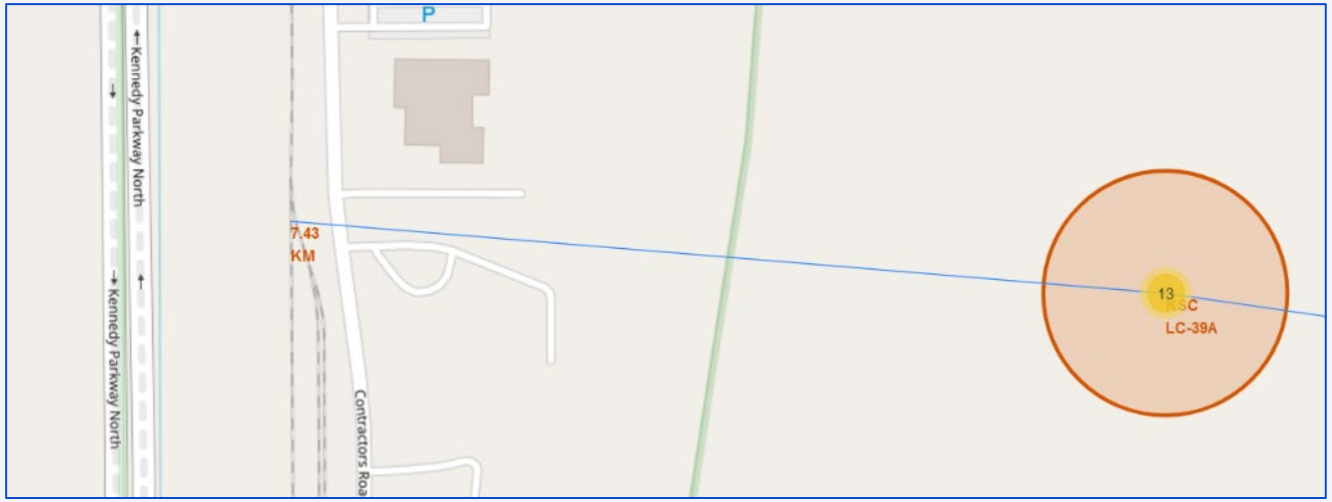
- All Launch Sites are near a coastline

Folium Markers to Marker Cluster for each Launch



Launch Site to Nearest Railway

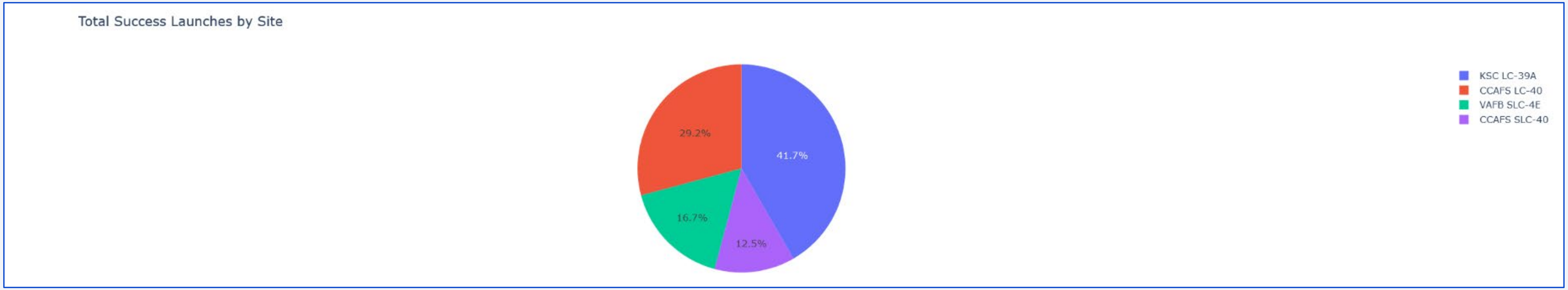
- Distance between
  - KSC LC-39A and
  - Nearest Railway
  - 7.43km



Section 4

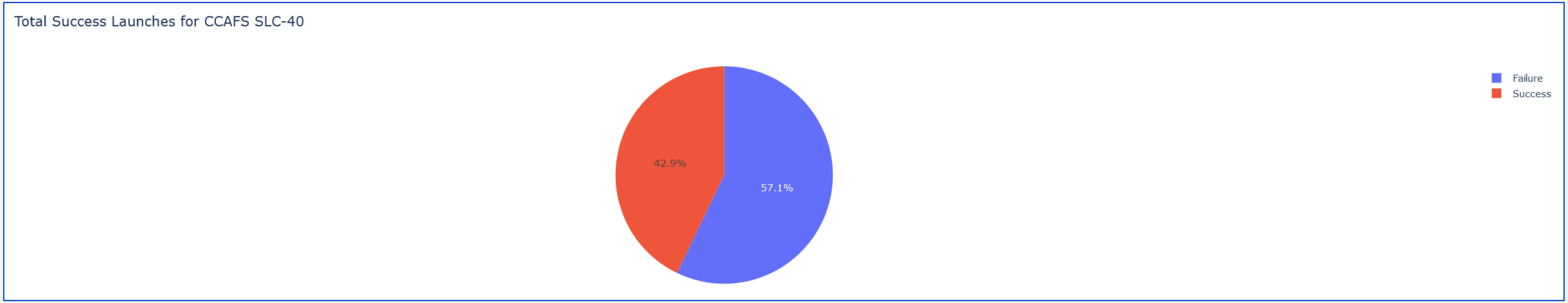
Build a Dashboard with Plotly Dash

Total Successful Launches by Launch Site

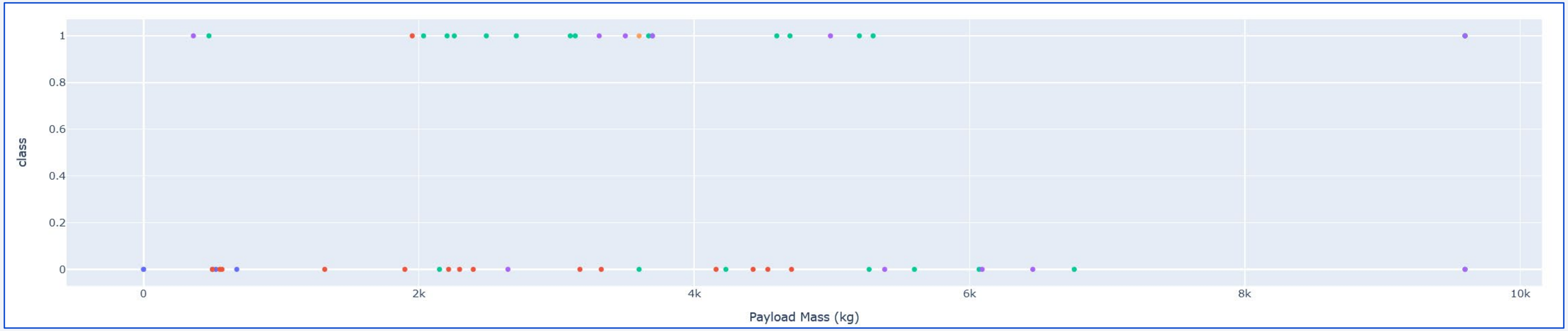


- KSC LC-39A has the most successful launches
- While CCAFS SLC-40 has the least

Total Successful Launches for CCAFS SLC-40



Payload Mass vs. Launch Outcome

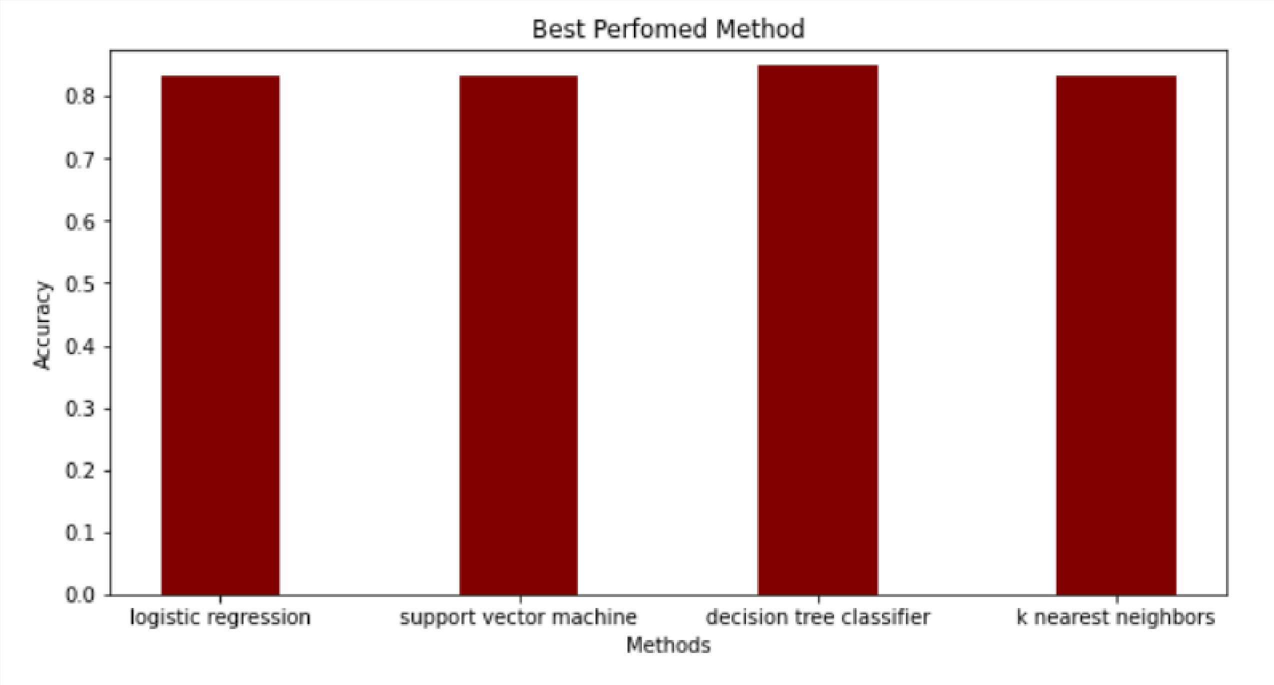


Section 5

Predictive Analysis (Classification)

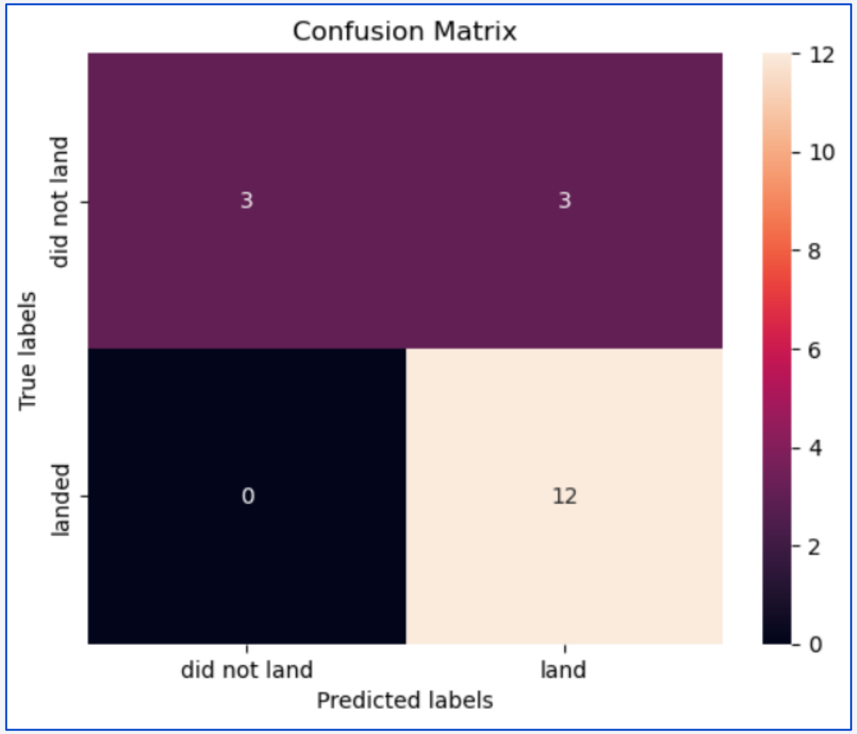
Classification Accuracy

- Bar Chart for model accuracy of all built classification models
  - Decision Tree Classifier has the highest accuracy



Confusion Matrix

- Show the confusion matrix of the Decision Tree Classifier
  - 12 True Negatives and 3 True Positives
    - Vs.
  - 0 False Negatives and 3 False Positives



Thank you!