IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Atharva Donde

Outline

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

Executive Summary

- Data was collected from public SpaceX API and SpaceX Wikipedia page.
- Explored the data using SQL, visualization, folium maps, and dashboards.
- Changed all categorical variables to binary using one hot encoding.
- Four machine learning models were produced: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors.
- Standardized data and used GridSearchCV to find best parameters for machine learning models.
- All models over predicted successful landings with similar accuracy rate.

Introduction

Background

- Space Y is founded by Billionaire industrialist Allon Musk.
- Unlike other rocket providers, SpaceX's Falcon 9 Can recover the first stage.
- This leads to SpaceX saving a lot on First stage.

Problem

- Space Y would like to compete with SpaceX, but due to SpaceX being able to recover Stage 1 it’s easier for them to manage cost.

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was processed
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

Data collection process involved a combination of API requests from Space X public API and web scraping data from a table in Space X's Wikipedia entry.

:Space X API Data Columns

FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

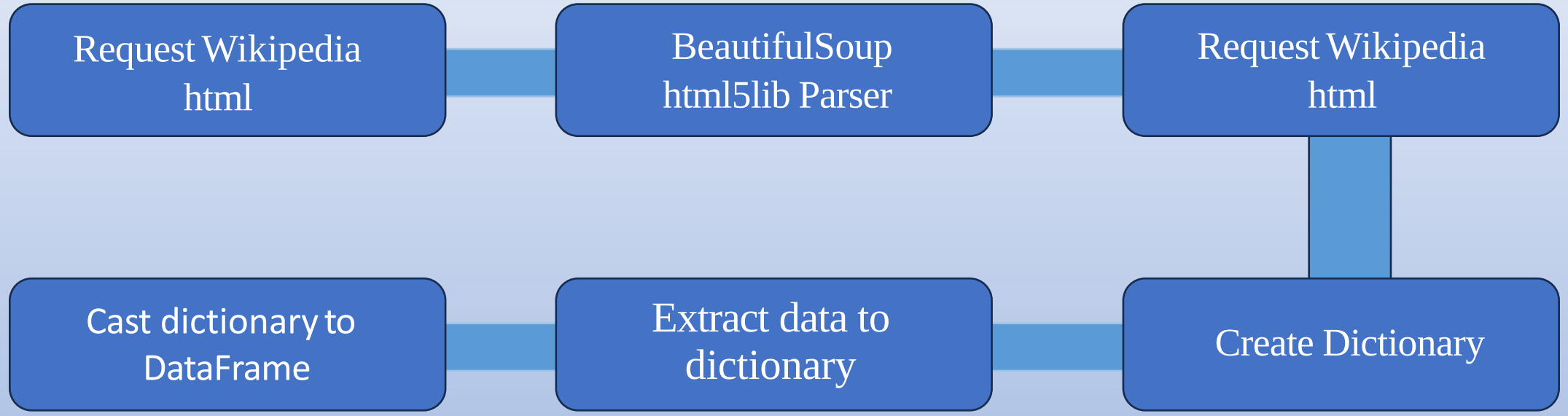
:Space X API Data Columns

Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection – SpaceX API

Github url: https://github.com/Atharva-612/Data-Science-Project/blob/main/data-collection-api.ipynb

Data Collection - Scraping



Github url: https://github.com/Atharva-612/Data-Science-Project/blob/main/Data%20wrangling.ipynb

Data Wrangling

- Create a training label with landing outcomes where successful = 1 & failure = 0.
- Outcome column has two components: ‘Mission Outcome’ ‘Landing Location’.
- New training label column ‘class’ with a value of 1 if ‘Mission Outcome’ is True and 0 otherwise.
- True ASDS, True RTLS, & True Ocean – set to -> 1.
- None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0.

Github url: https://github.com/Atharva-612/Data-Science-Project/blob/main/Data%20wrangling.ipynb

EDA with Data Visualization

- Exploratory Data Analysis performed on variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.

- Scatter plots, line charts, and bar plots were used to compare relationships between variables to decide if a relationship exists so that they could be used in training the machine learning model.

- Plots used: Flight Number vs. Payload Mass, Flight Number vs. Launch Site, Payload Mass vs. Launch Site, Orbit vs. Success Rate, Flight Number vs. Orbit, Payload vs Orbit, and Success Yearly Trend.

Github url: https://github.com/Atharva-612/Data-Science-Project/blob/main/Exploratory%20Data%20Analysis%20with%20visualization.ipynb

EDA with SQL

- Loaded data set into IBM DB2 Database.
- Queried using SQL Python integration.
- Queries were made to get a better understanding of the dataset.
- Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes.

Github URL: https://github.com/Atharva-612/Data-Science-Project/blob/main/Explaratory%20Data%20Analysis%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast, and City.

- This allows us to understand why launch sites may be located where they are. Also visualizes successful landings relative to location.

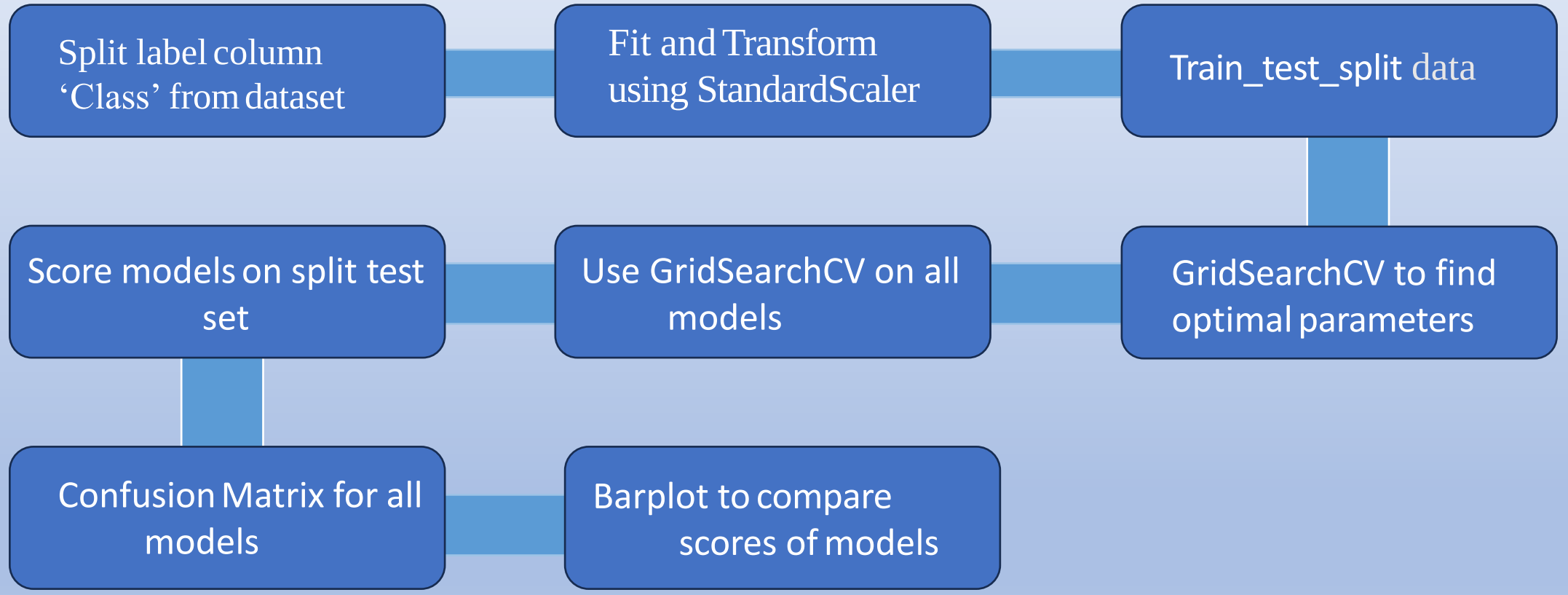
Github URL: https://github.com/Atharva-612/Data-Science-Project/blob/main/Interactive%20visual%20analytics%20with%20folium.ipynb

Build a Dashboard with Plotly Dash

- Dashboard includes a pie chart and a scatter plot.
- Pie chart can be selected to show distribution of successful landings across all launch sites and can be selected to show individual launch site success rates.
- Scatter plot takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.
- The pie chart is used to visualize launch site success rate.
- The scatter plot can help us see how success varies across launch sites, payload mass, and booster version category.

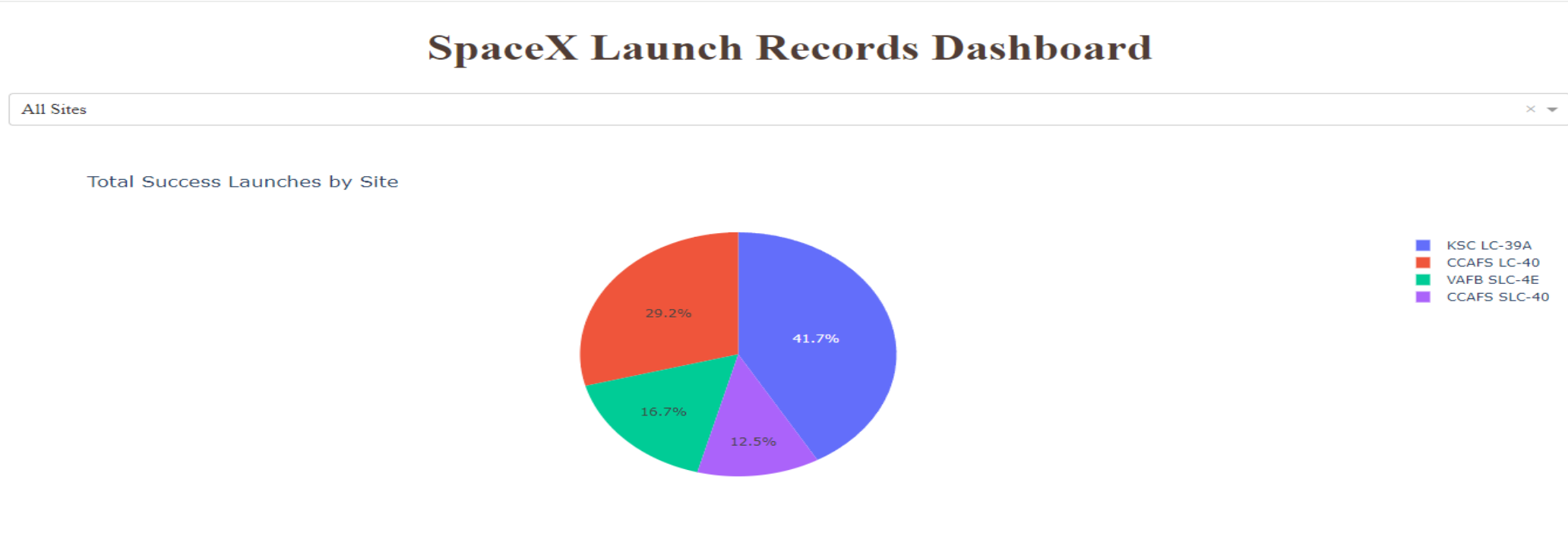
Github URL: https://github.com/Atharva-612/Data-Science-Project/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)



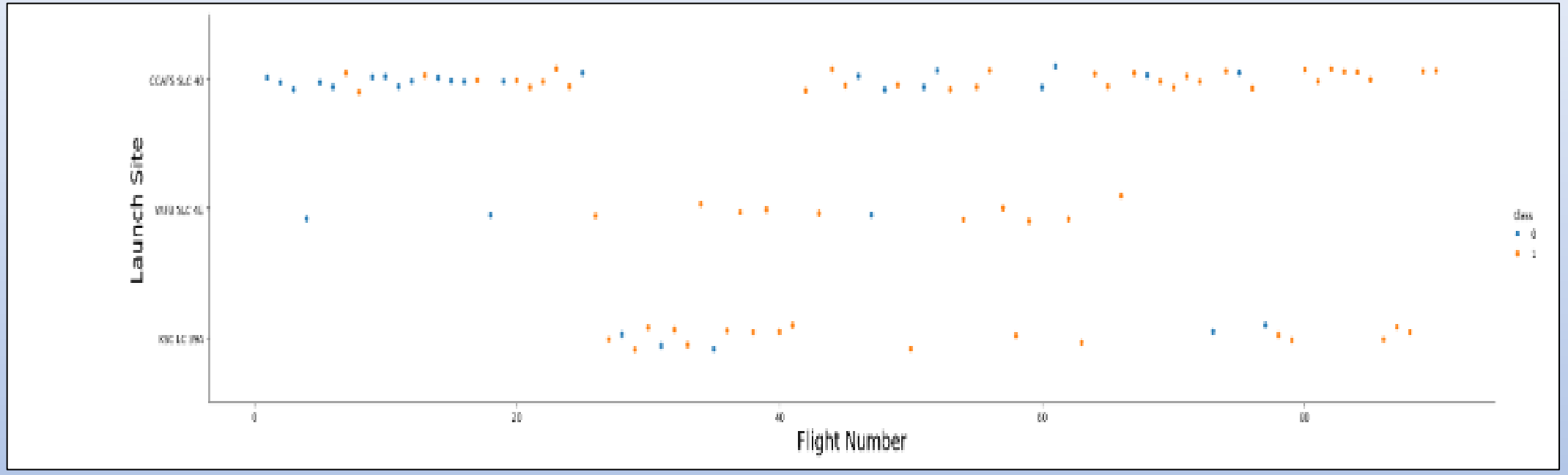
Github url:https://github.com/Atharva-612/Data-Science-Project/blob/main/Predictive%20analysis.ipynb

Results



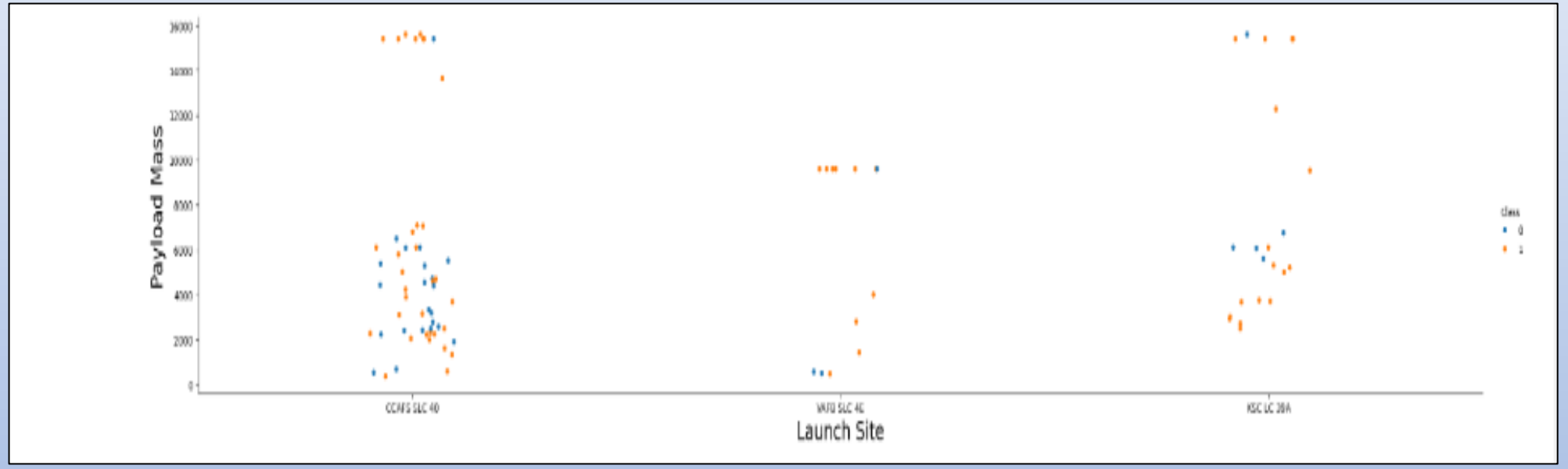
The following shows the results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and finally the results of our model with about 83% accuracy.

Flight Number vs. Launch Site



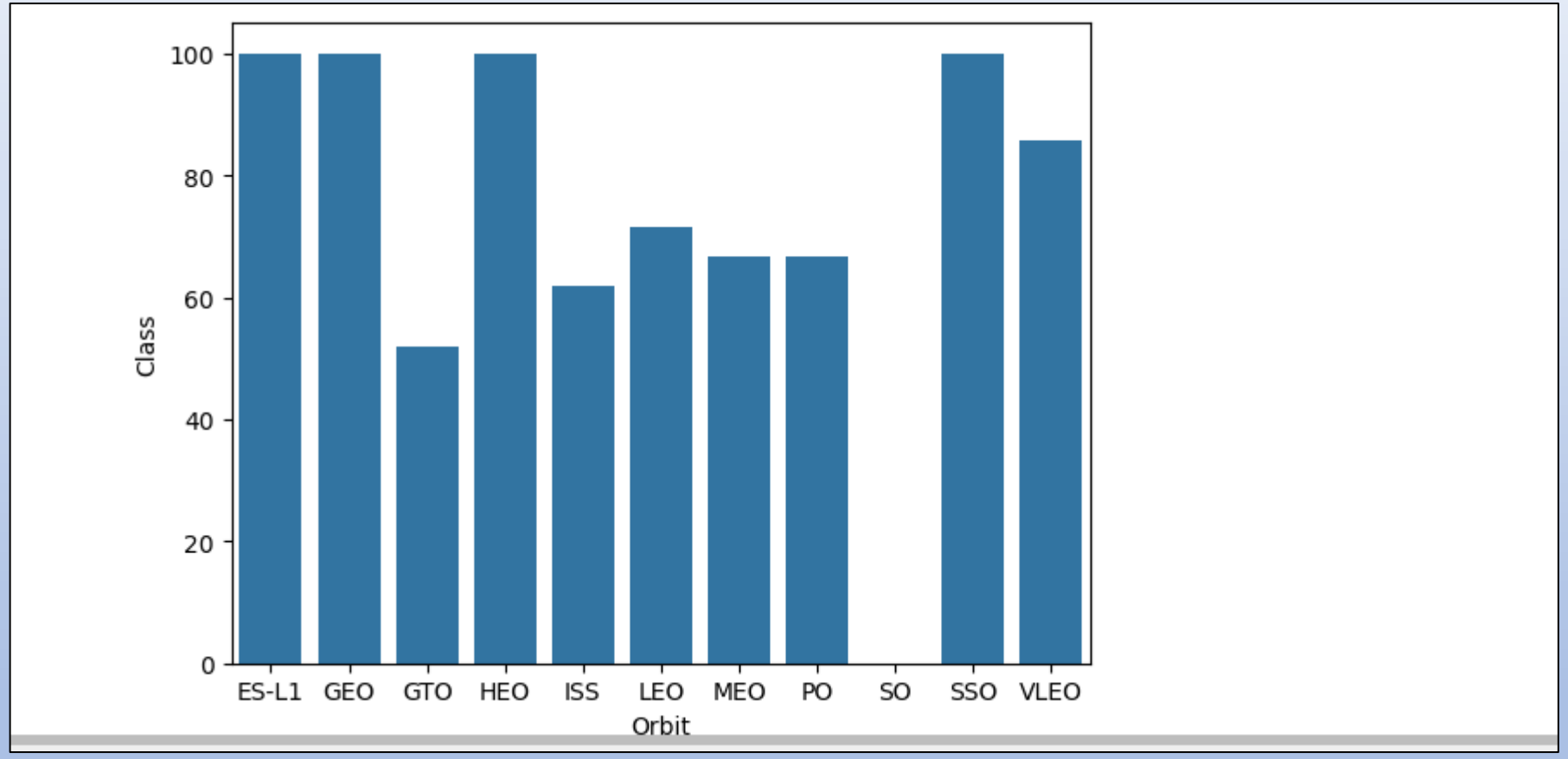
- Blue indicates unsuccessful launch and orange indicates successful launch.
- This graph indicates that launch success rate has increased over time.

Payload vs. Launch Site



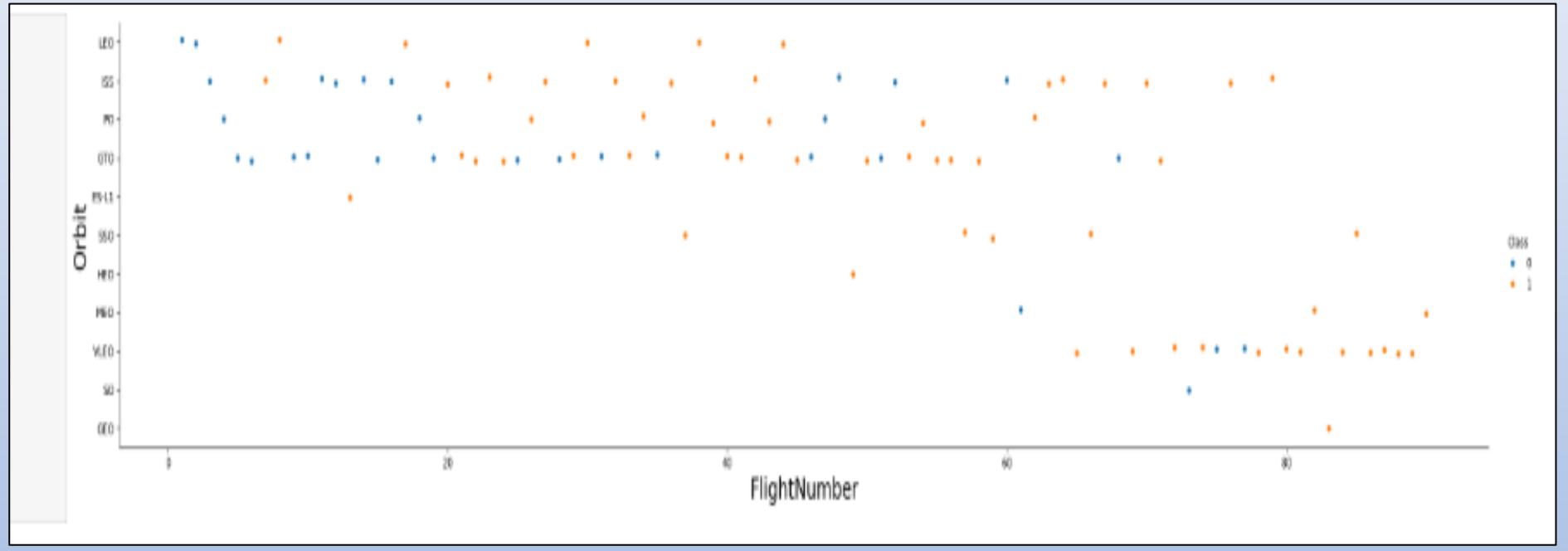
- Payload mass varies with Launch site.
- Payload mass appears to fall mostly between 0-6000 kg.

Success Rate vs. Orbit Type



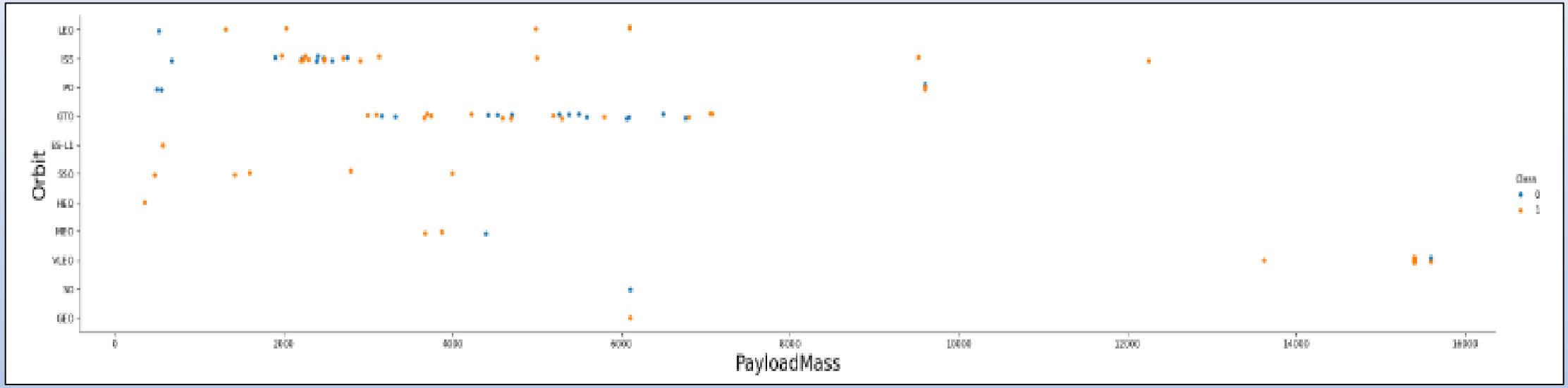
- SS0, GEO, ES-L1 receives 100% success rate.
- SO has a success rate of 0%.

Flight Number vs. Orbit Type



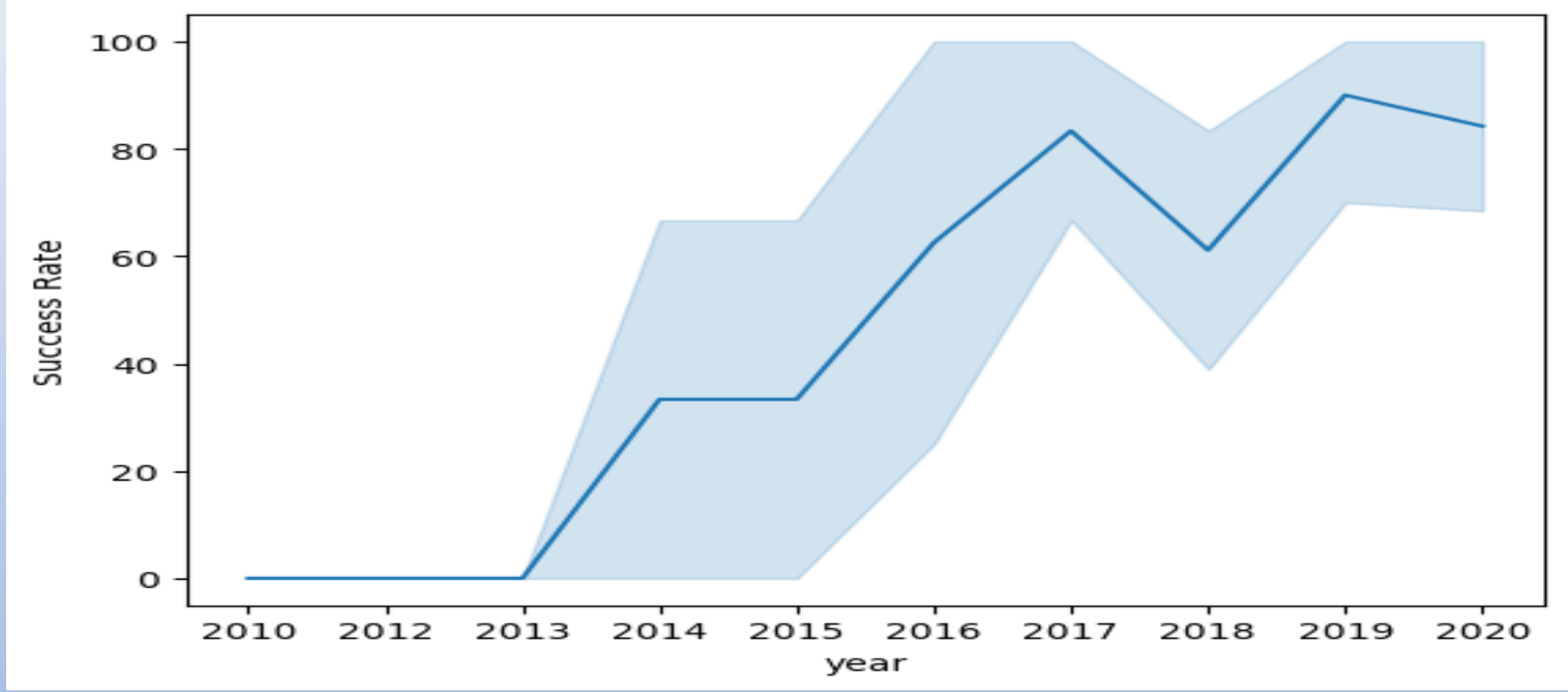
- Launch Orbit preferences changed over Flight Number.
- SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches.

Payload vs. Orbit Type



- Payload mass seems to correlate with orbit.
- LEO and SSO seem to have relatively low payload mass.

Launch Success Yearly Trend



- Success generally increases over time since 2013 with a slight dip in 2018.
- Success rate is nearly 80%.

All Launch Site Names

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

- Query unique launch site names from database.
- CCAFS SLC-40 and CCAFSSLC-40 likely all represent the same launch site with data entry errors .
- CCAFS LC-40 was the previous name.
- Likely only 3 unique launch_site values: CCAFS SLC-40, KSC LC-39A, VAFB SLC-4E

Launch Site Names Begin with 'CCA'

```
[12]: %sql SELECT * FROM SPACEXTABLE WHERE Launch_Site LIKE 'CCA%' LIMIT 5
```

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

[12]:

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

- These are the first 5 entries that begin with ‘CCA’

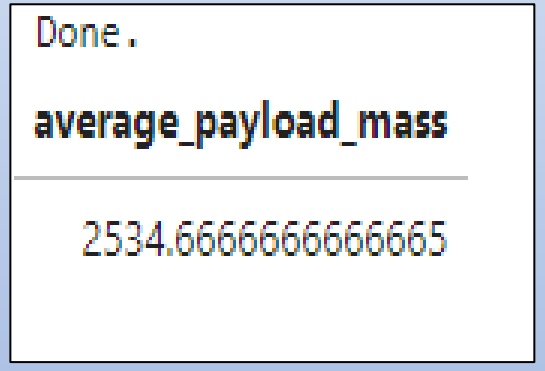
Total Payload Mass

Done.

SUM(PAYLOAD_MASS__KG_)
45596

- This query sums the total payload mass in kg where NASA was the customer.
- CRS stands for Commercial Resupply Services which indicates that these payloads were sent to the International Space Station (ISS).

Average Payload Mass by F9 v1.1



- This query calculates the average payload mass or launches which used booster version F9 v1.1.
- Average payload mass of F9 1.1 is on the low end of our payload mass range

First Successful Ground Landing Date

Date
2010-06-04

- This query returns the first successful ground pad landing date.
- First ground pad landing wasn't until the mid 2010.

Successful Drone Ship Landing with Payload between 4000 and 6000

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

This query returns the four booster versions that had successful drone ship landings and a payload mass between 4000 and 6000.

Total Number of Successful and Failure Mission Outcomes

Done.

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

- SpaceX appears to achieve its mission outcome nearly 99% of the time.
- This means that most of the landing failures are intended.
- Interestingly, one launch has an unclear payload status and unfortunately one failed in flight.

Boosters Carried Maximum Payload

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

- This query returns the booster versions that carried the highest payload mass of 15600 kg.
- These booster versions are very similar and all are of the F9 B5 B10xx.x variety.
- This likely indicates payload mass correlates with the booster version that is used.

2015 Launch Records

Done.

MONTH	landing__outcome	booster_version	payload_mass__kg_	launch_site
January	Failure (drone ship)	F9 v1.1 B1012	2395	CCAFS LC-40
April	Failure (drone ship)	F9 v1.1 B1015	1898	CCAFS LC-40

This query returns the Month, Landing Outcome, Booster Version, Payload Mass (kg), and Launch site of 2015 launches where stage 1 failed to land on a drone ship.

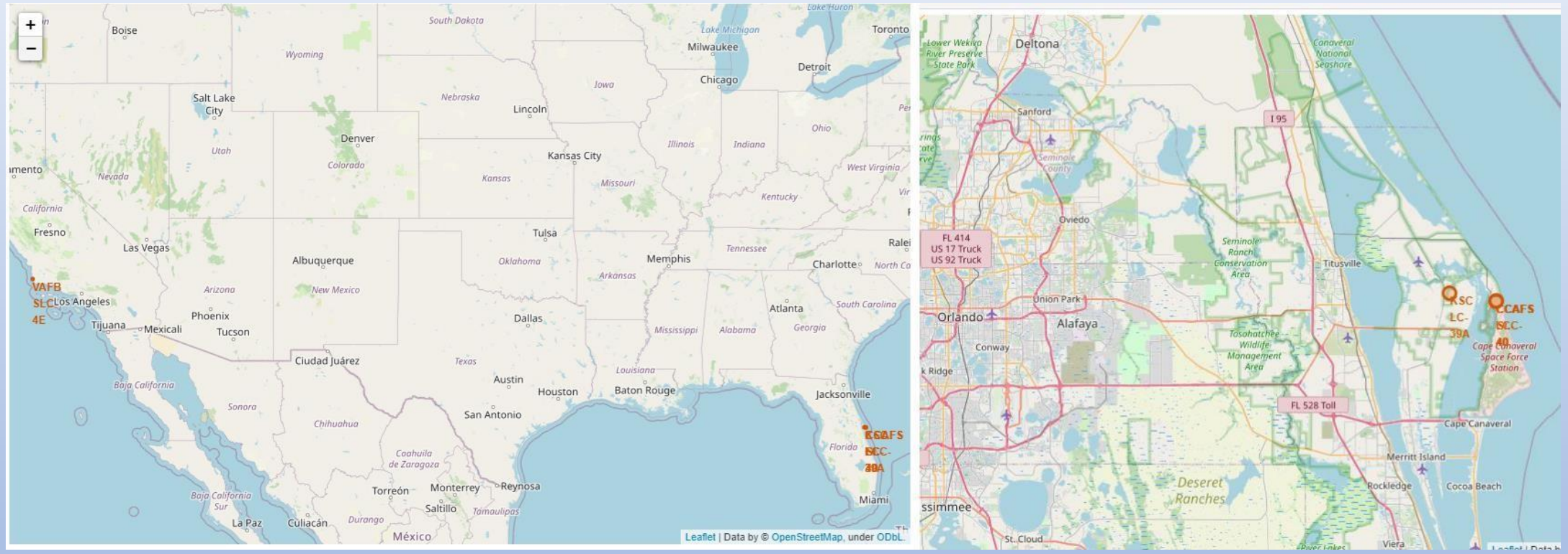
Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

Done.

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

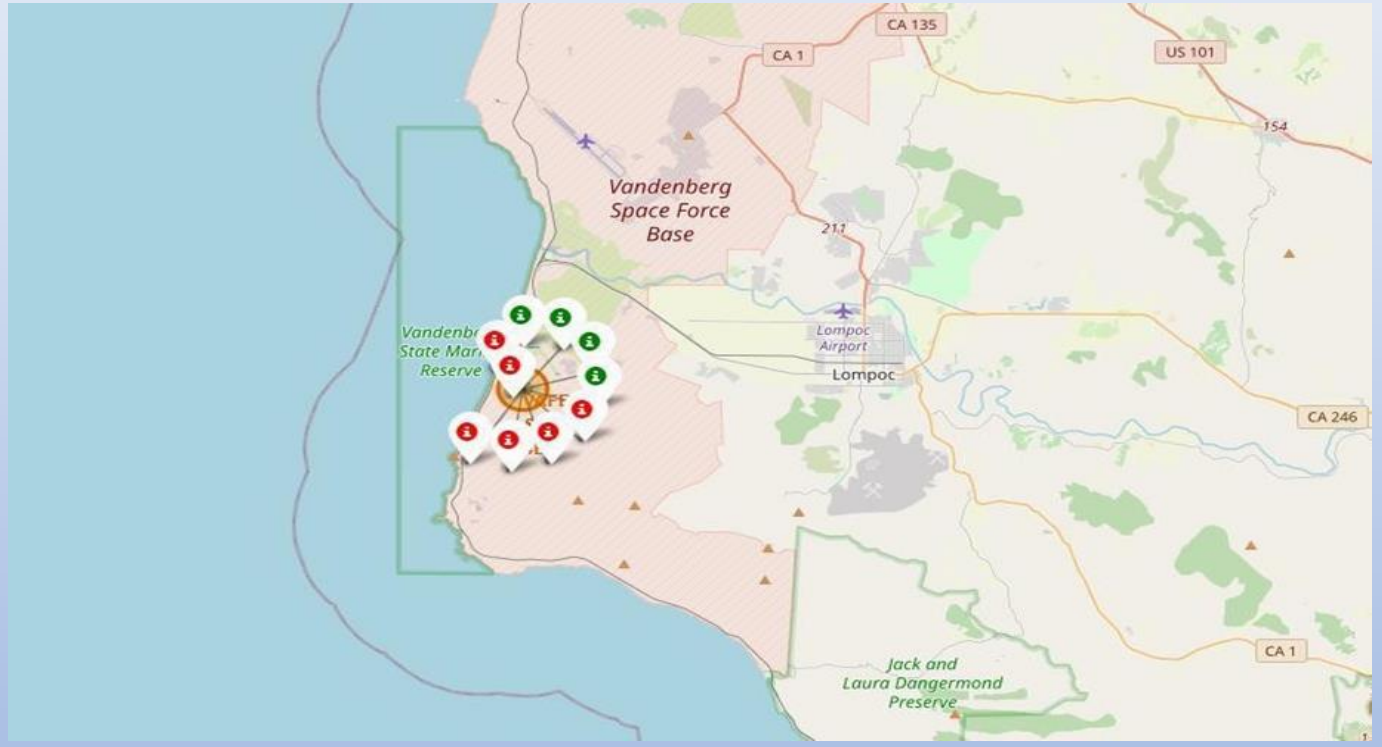
- This query returns a list of successful landings and between 2010-06-04 and 2017-03-20 inclusively.
- There are two types of successful landing outcomes: drone ship and ground pad landings.
- There were 8 successful landings in total during this time period.

Launch Site Locations



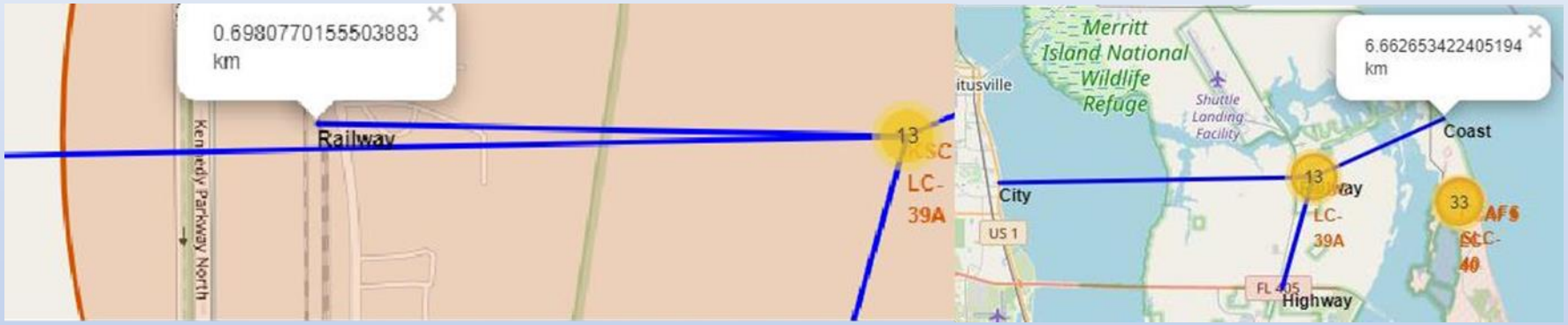
The left map shows all launch sites relative US map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

Color-Coded Launch Markers



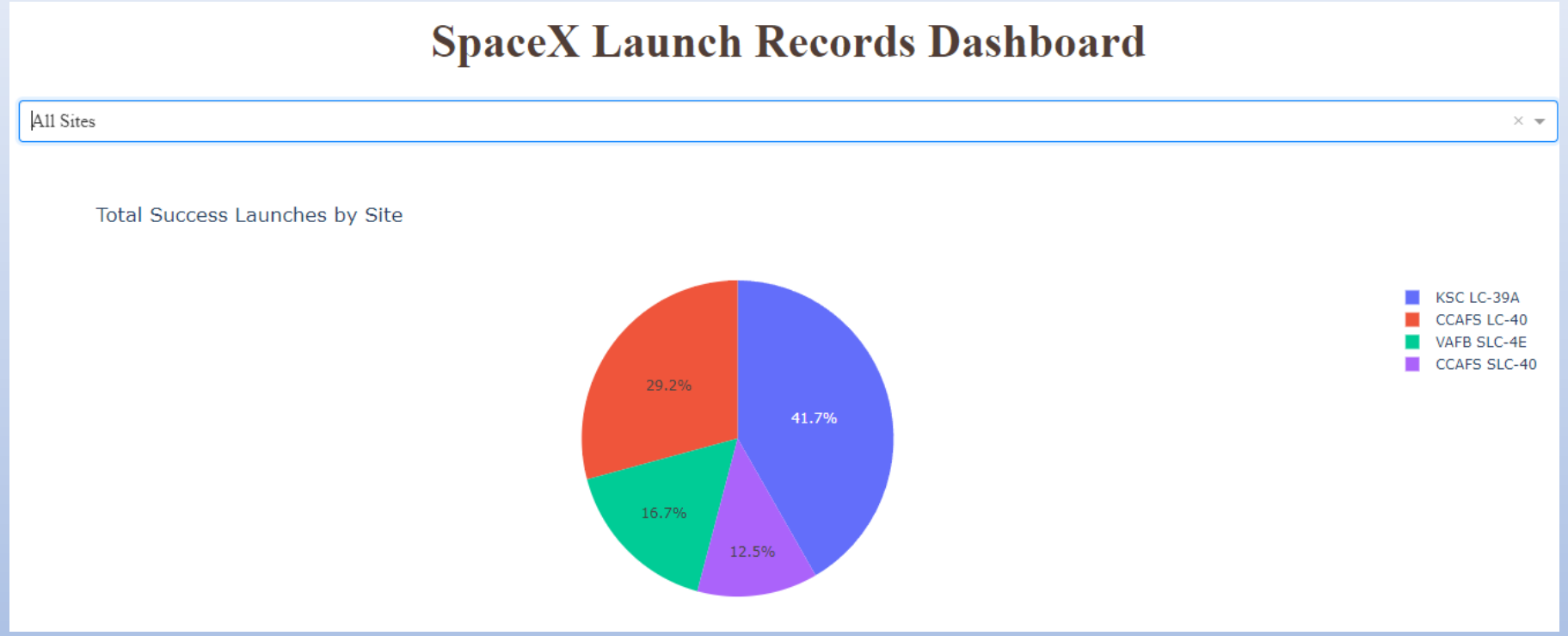
- Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon). In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings.

Location Proximities



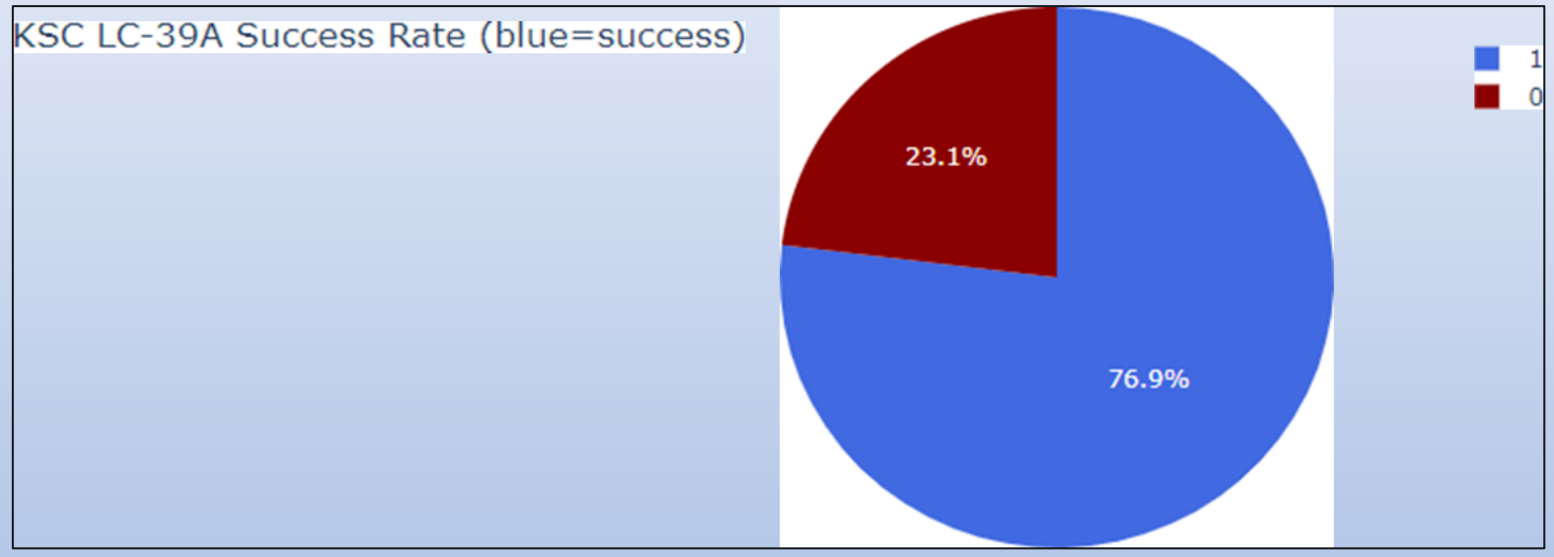
- Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas.

Successful Launches Across Launch Sites



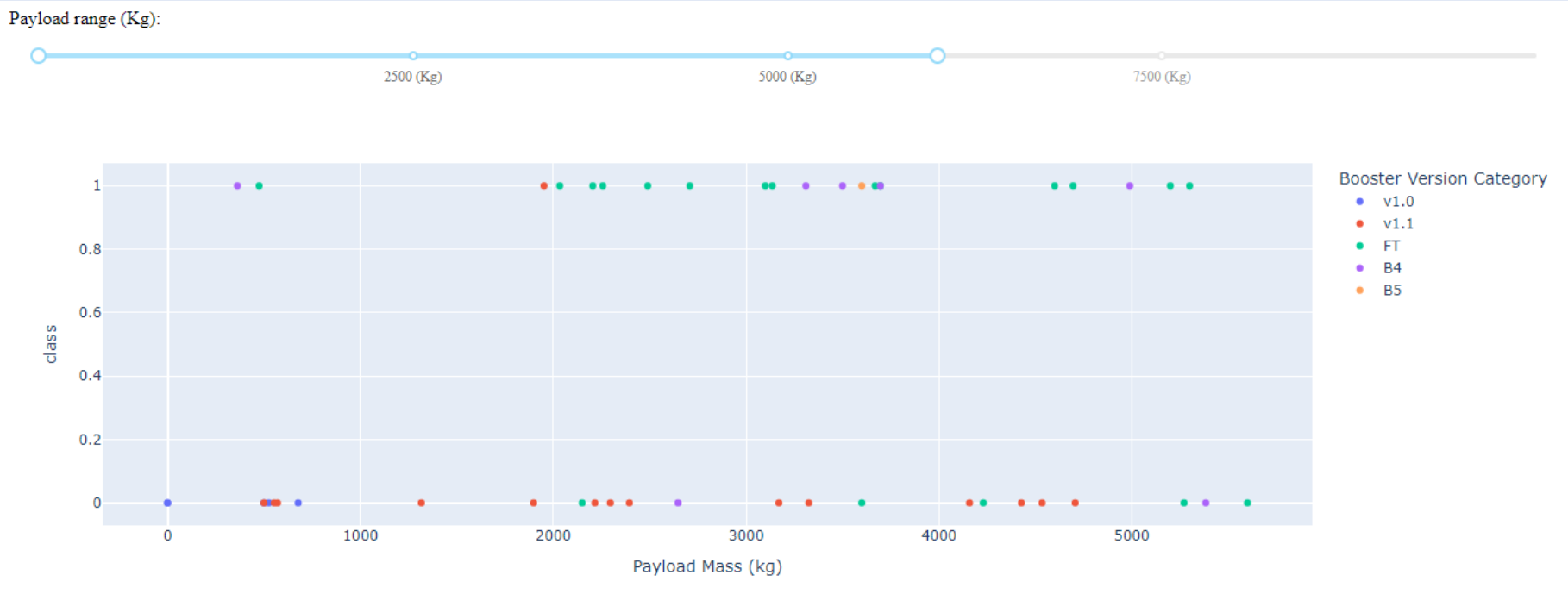
CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings where performed before the name change. VAFB has the smallest share of successful landings. This may be due to smaller sample and increase in difficulty of launching in the west coast.

Highest Success Rate Launch Site



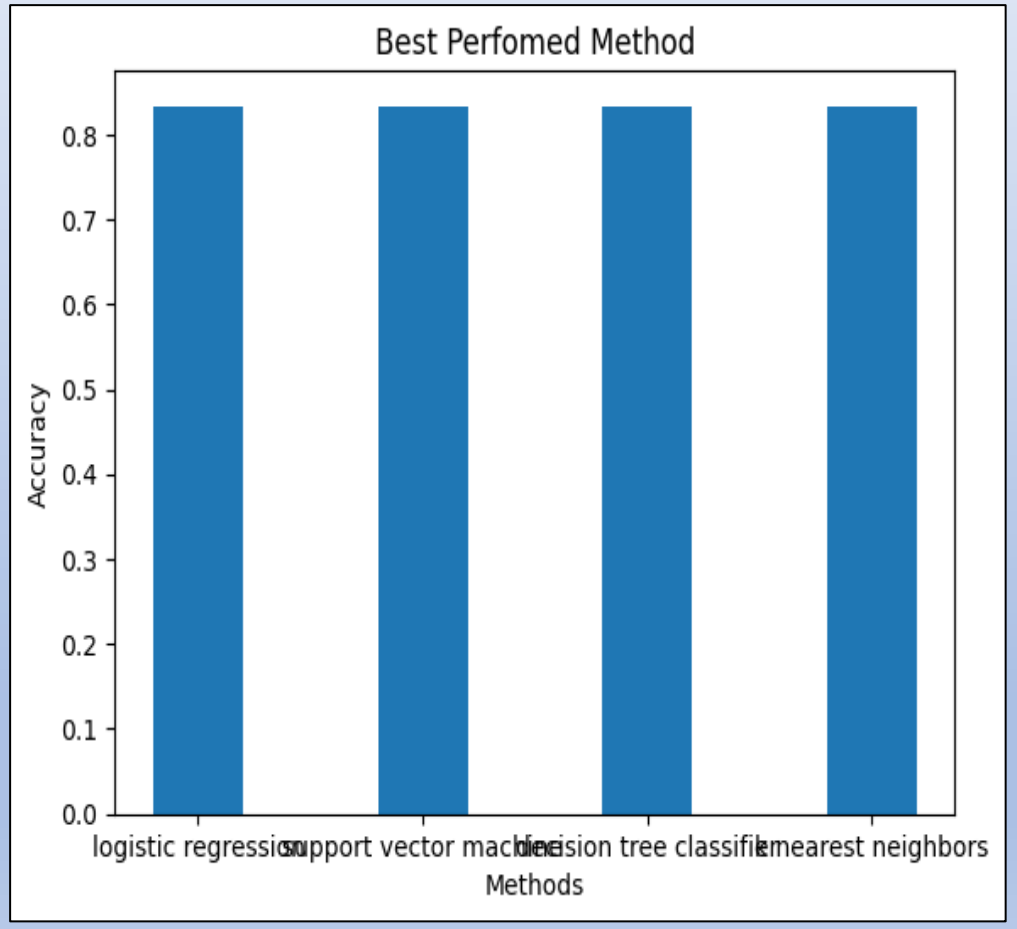
KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings.

<Dashboard Screenshot 3>



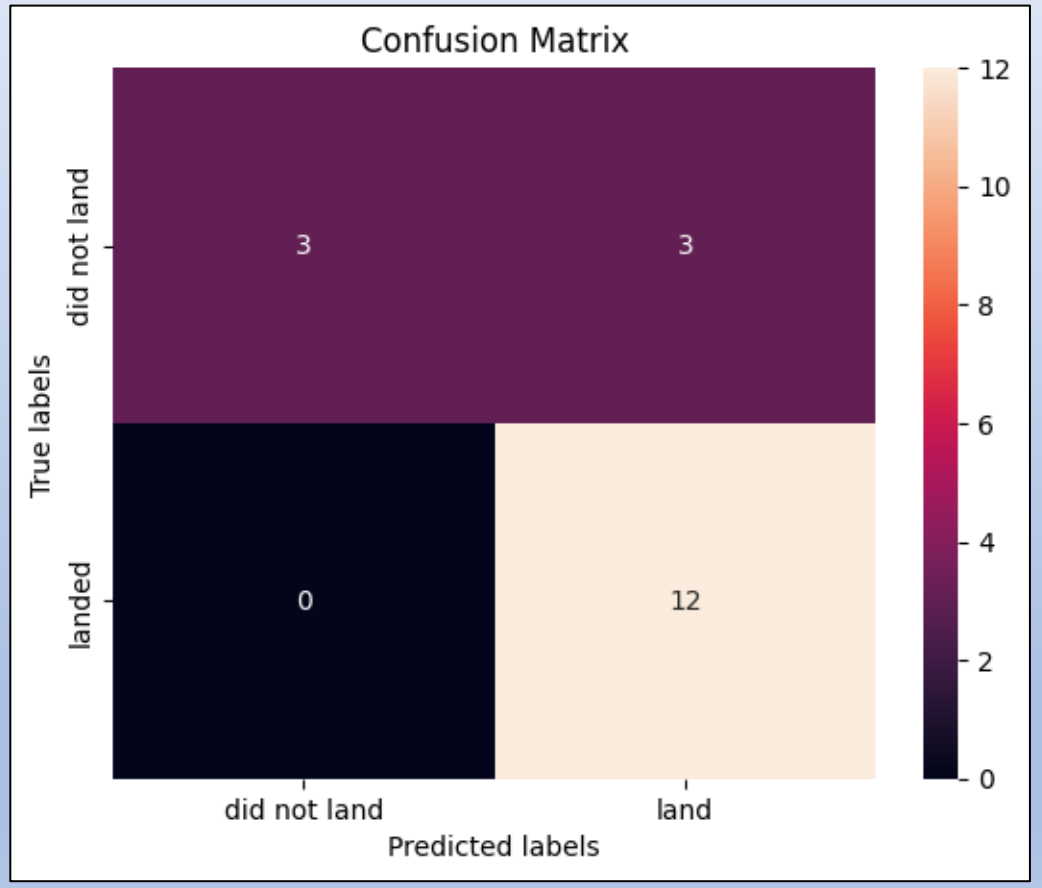
Plotly dashboard has a Payload range selector. Scatter plot accounts for booster version category in color and number of launches in point size. In this particular range of 0-6000, interestingly there are two failed landings with payloads of zero kg.

Classification Accuracy



- All models had virtually the same accuracy on the test set at 83.33% accuracy. It should be noted that test size is small at only sample size of 18.
- This can cause large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.
- We likely need more data to determine the best model.

Confusion Matrix



- The models predicted 12 successful landings when the true label was successful landing.
- The models predicted 3 unsuccessful landings when the true label was unsuccessful landing.
- The models predicted 3 successful landings when the true label was unsuccessful landings (false positives).
- Our models over predict successful landings.

Conclusions

- Our task: to develop a machine learning model for Space Y who wants to bid against SpaceX.
- The goal of model is to predict when Stage 1 will successfully land to save ~$100 million USD.
- Used data from a public SpaceX API and web scraping SpaceX Wikipedia page.
- Created data labels and stored data into a DB2 SQL database.
- Created a dashboard for visualization.
- We created a machine learning model with an accuracy of 83%.
- Allon Mask of SpaceY can use this model to predict with relatively high accuracy whether a launch will have a successful Stage 1 landing before launch to determine whether the launch should be made or not.

Appendix

GitHub repository url: https://github.com/Atharva-612/Data-Science-Project