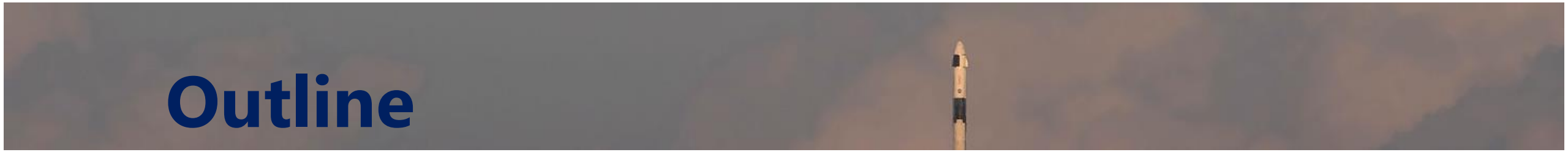
Data Science Capstone Project

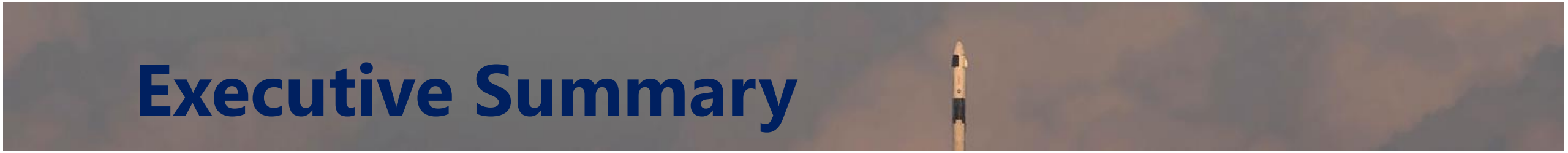
Stephen Hrithick

March 15th , 2025

Outline



Executive Summary



The aim of this research is to analyze Falcon 9 data from SpaceX gathered through various sources and apply machine learning models to predict the success of the first-stage landing. This will enable other space agencies to assess whether they should compete with SpaceX in bidding opportunities.

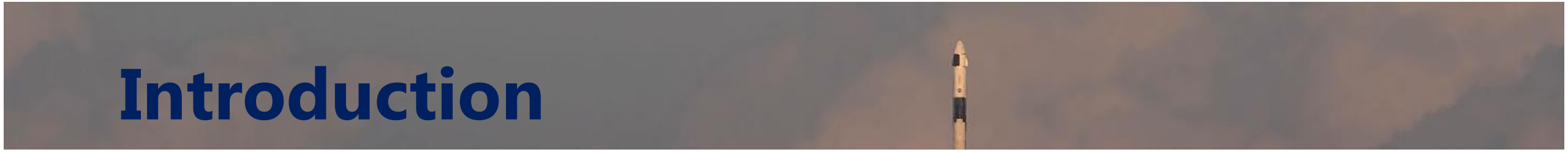
The following methods were used to collect, analyze, and model the data, as well as to make predictions:

- **Data Collection**: Gathered data via API and web scraping.
- **Data Transformation**: Applied data wrangling techniques to clean and structure the data.
- **Exploratory Data Analysis**: Used SQL and data visualizations to explore the dataset.
- **Interactive Map**: Created a map with Folium to examine the proximity of launch sites.
- **Dashboard Development**: Built an interactive dashboard with Plotly Dash to analyze launch records.
- **Predictive Modeling**: Developed a machine learning model to predict the success of Falcon 9's first-stage landing.

Summary of all results:

- Exploratory Data Analysis results.
- Interactive analytics demo in screenshots.
- Predictive analysis results

Introduction

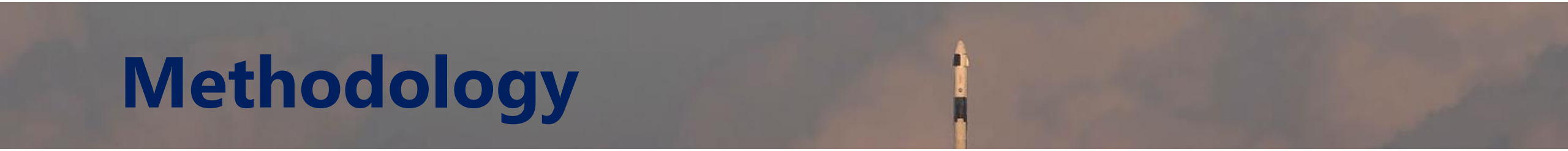


Project background and context

- SpaceX has emerged as the leading company in the commercial space industry, revolutionizing space travel by making it more affordable. The company advertises Falcon 9 launches at a cost of $62 million, while competitors charge upwards of $165 million per launch. A significant portion of the savings comes from SpaceX's ability to reuse the first stage of the rocket. By predicting whether the first stage will successfully land, we can estimate the overall cost of a launch. Using public data and machine learning models, we aim to predict whether SpaceX will be able to reuse the first stage.

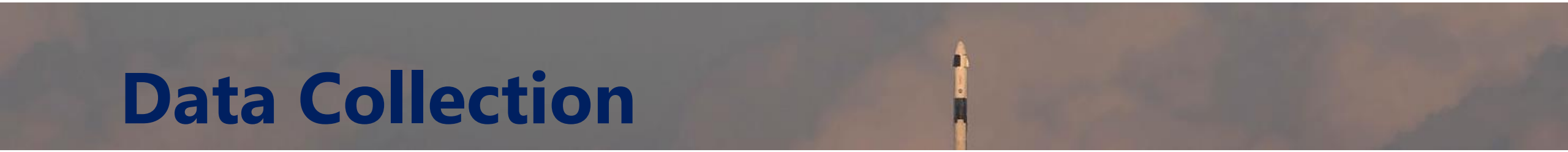
- **Problem**
  - How do variables such as payload mass, launch site, number of flights, and orbits affect the success of the first stage landing?
  - Does the rate of successful landings increase over the years?
  - What is the best algorithm that can be used for binary classification in this case?

Methodology



- Data collection methodology:
  - Combined data from SpaceX public API and SpaceX Wikipedia page
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Tuned models using GridSearchCV

Data Collection



- Data collection involves gathering information from various available sources, which can be structured, unstructured, or semi-structured. For this project, data was obtained through the SpaceX API and web scraping of relevant launch data from Wikipedia pages.
- **Data Columns are obtained by using SpaceX REST API:** FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude.
- **Data Columns are obtained by using Wikipedia Web Scraping:** Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date, Time

Data Collection

Data Collection – SpaceX API

Request (SpaceX APIs)

↓

.JSON file + Lists(Launch Site, Booster Version, Payload Data)

↓

Json_normalize to DataFrame data from JSON

↓

Filter data to only include Falcon 9 launches

↓

Cast dictionary to a DataFrame

↓

Dictionary relevant data

↓

Imputate missing PayloadMass values with mean

Data Collection – Web Scraping

Request Wikipedia html

↓

BeautifulSoup html5lib Parser

↓

Find launch info html table

↓

Create dictionary

↓

Iterate through table cells to extract data to dictionary

↓

Cast dictionary to DataFrame

Data Wrangling

- The dataset includes various instances where the booster did not land successfully. In some cases, a landing attempt failed due to an accident. For example, "True Ocean" indicates a successful landing in a designated ocean region, while "False Ocean" signifies an unsuccessful landing in the same region. Similarly, "True RTLS" represents a successful landing on a ground pad, while "False RTLS" indicates a failed ground pad landing.
- "True ASDS" refers to a successful landing on a drone ship, and "False ASDS" denotes an unsuccessful landing on a drone ship. These outcomes are primarily converted into training labels, where "1" signifies a successful landing and "0" indicates an unsuccessful one.

Perform exploratory Data Analysis and determine Training Labels

Calculate the number of launches on each site

Calculate the number and occurrence of each orbit

Calculate the number and occurrence of mission outcome per orbit type

Create a landing outcome label from Outcome column

Exporting the data to CSV

https://github.com/Stephen-507/Python-/blob/6037a1e2f5073354bb9a0725e840d30640801f8a/Data%20Wrangling.ipynb

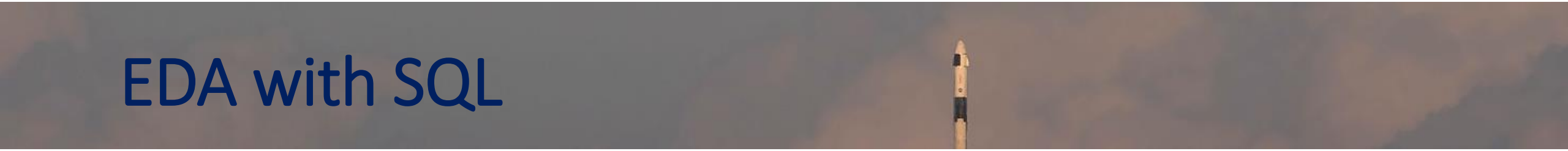
EDA with Data Visualization

Exploratory Data Analysis (EDA) was conducted on the variables: Flight Number, Payload Mass, Launch Site, Orbit, Class, and Year. The following visualizations were created:

- 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
- Yearly Success Trend

https://github.com/Stephen-507/Python-/blob/6037a1e2f5073354bb9a0725e840d30640801f8a/EDA%20with%20Data%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

https://github.com/Stephen-507/Python-/blob/6037a1e2f5073354bb9a0725e840d30640801f8a/EDA%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.

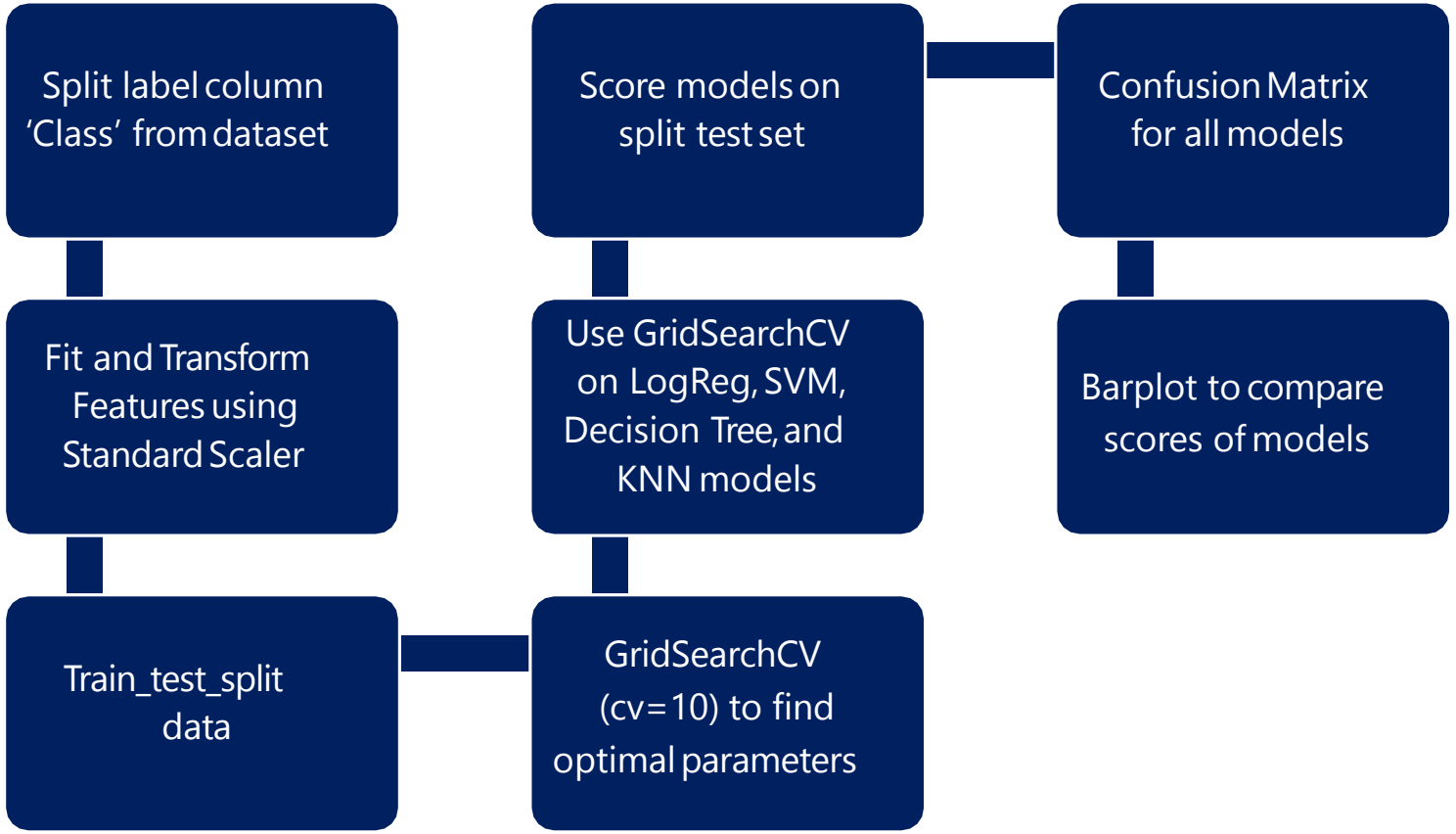
https://github.com/Stephen-507/Python-/blob/6037a1e2f5073354bb9a0725e840d30640801f8a/Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- The dashboard features a pie chart and a scatter plot for data visualization.
- The pie chart can display the distribution of successful landings across all launch sites or show the success rate for individual launch sites.
- The scatter plot takes two inputs: it allows users to select either all sites or a specific site, and features a slider to adjust payload mass between 0 and 10,000 kg.
- The pie chart provides a clear view of the success rate for each launch site, while the scatter plot helps visualize how success rates vary based on launch site, payload mass, and booster version.

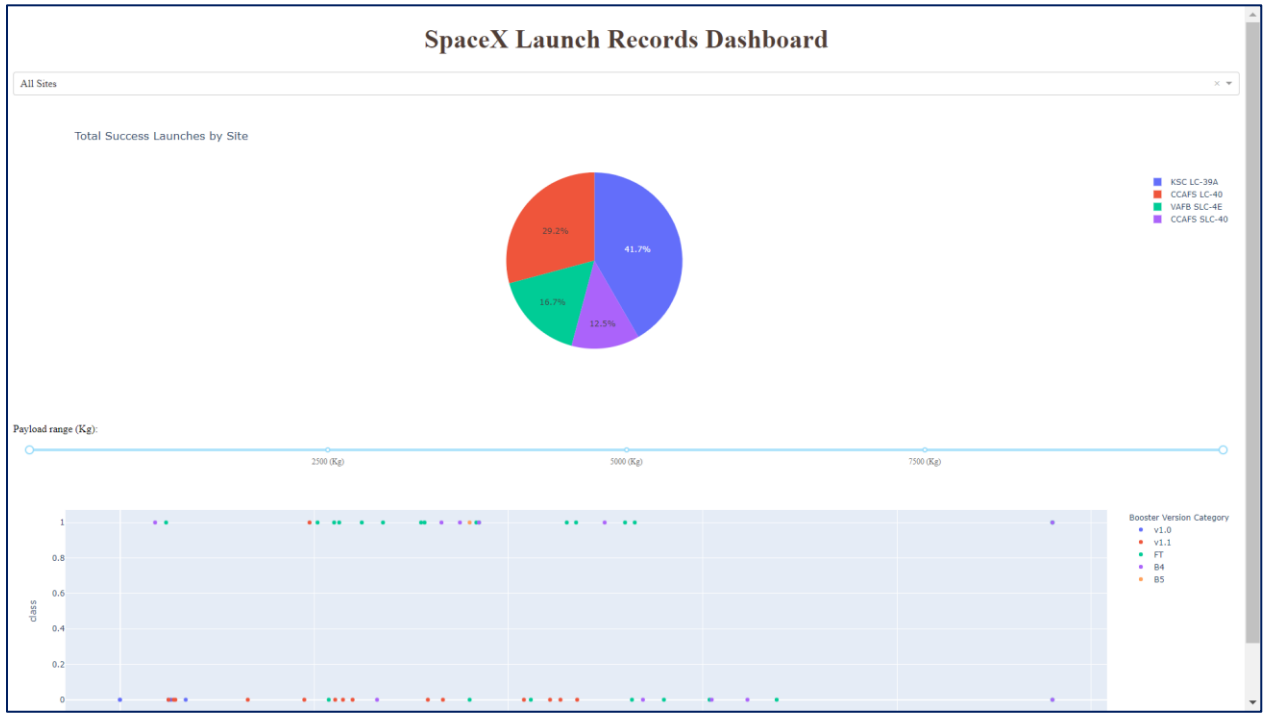
https://github.com/Stephen-507/Python-/blob/ee562efd0ab0420690214c7e43e2185bc0058bbb/spacex_dash_app.py

Predictive analysis (Classification)



https://github.com/Stephen-507/Python-/blob/ee562efd0ab0420690214c7e43e2185bc0058bbb/Machine%20Learning%20Prediction.ipynb

Results

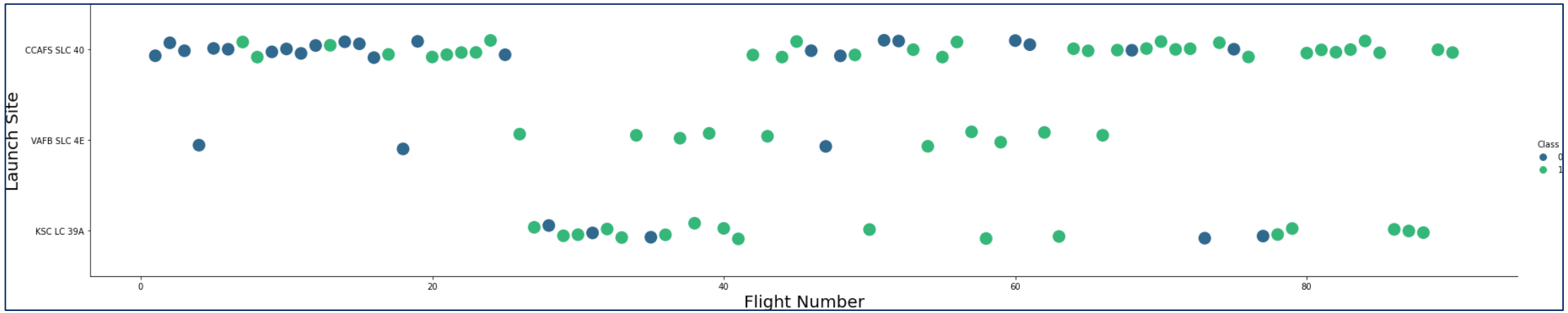


This is a preview of the Plotly dashboard. The following sides will show the results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and finally the results of our model with about 83% accuracy.

EDA WITH VISUALIZATION

EXPLORATORY DATA ANALYSIS WITH SEABORN PLOTS

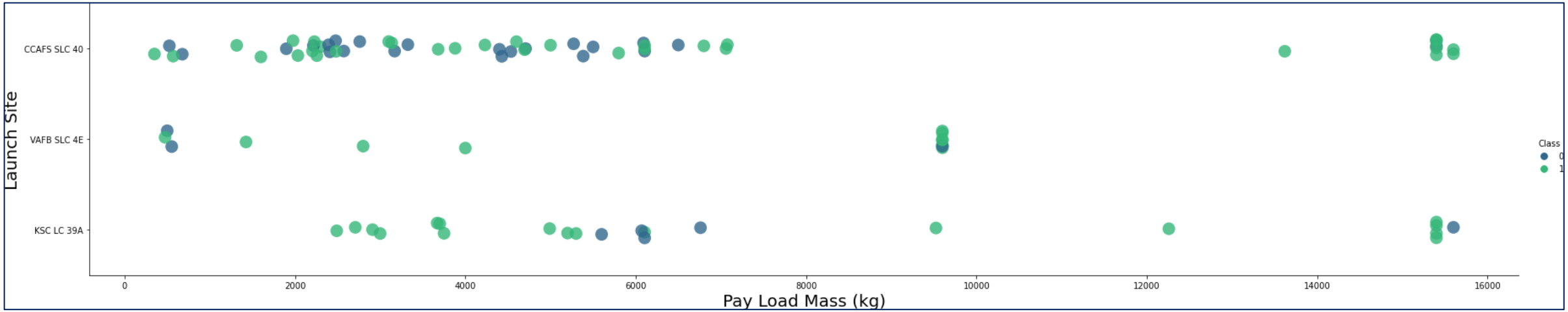
Flight Number vs. Launch Site



Green indicates successful launch; Purple indicates unsuccessful launch.

Graphic suggests an increase in success rate over time (indicated in Flight Number). Likely a big breakthrough around flight 20 which significantly increased success rate. CCAFS appears to be the main launch site as it has the most volume.

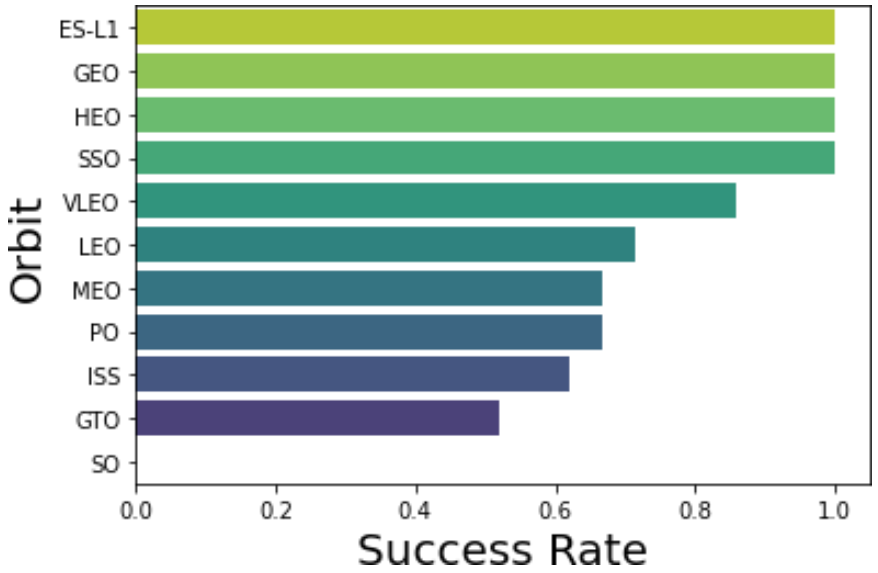
Payload vs. Launch Site



Green indicates successful launch; Purple indicates unsuccessful launch.

Payload mass appears to fall mostly between 0-6000 kg. Different launch sites also seem to use different payload mass.

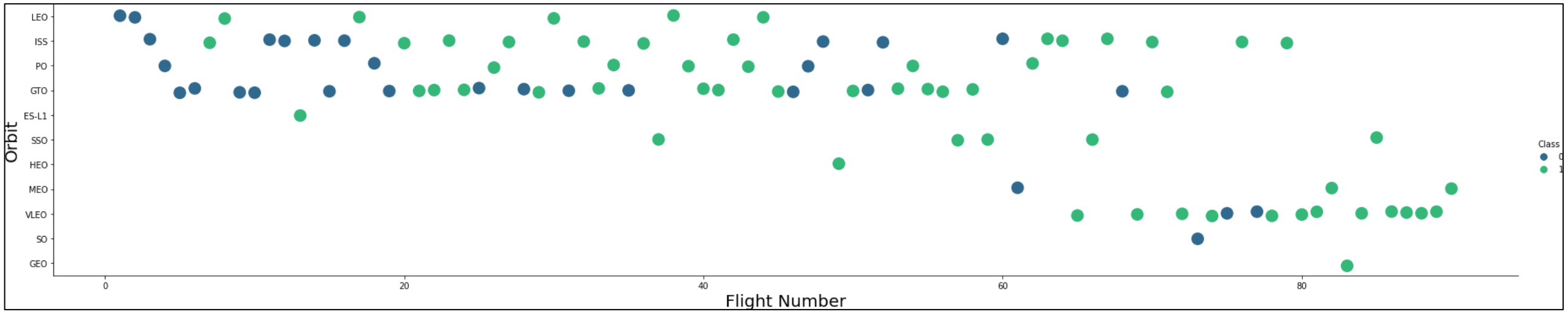
Success rate vs. Orbit type



Success Rate Scale:
0 as 0%
0.6 as 60%
1 as 100%

- ES-L1 (1), GEO (1), HEO (1) have 100% success rate (sample sizes in parenthesis) SSO (5) has 100% success rate
- VLEO (14) has decent success rate and attempts
- SO (1) has 0% success rate
- GTO (27) has the around 50% success rate but largest sample

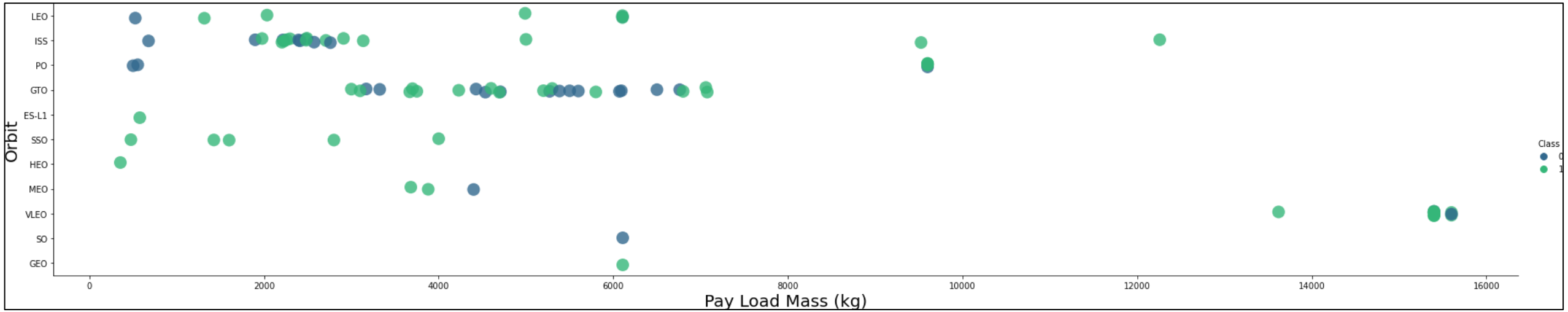
Flight Number vs. Orbit type



Green indicates successful launch; Purple indicates unsuccessful launch.

- Launch Orbit preferences changed over Flight Number.
- Launch Outcome seems to correlate with this preference.
- SpaceX started with LEO orbits which saw moderate success LEO and returned to VLEO in recent launches
- SpaceX appears to perform better in lower orbits or Sun-synchronous orbits

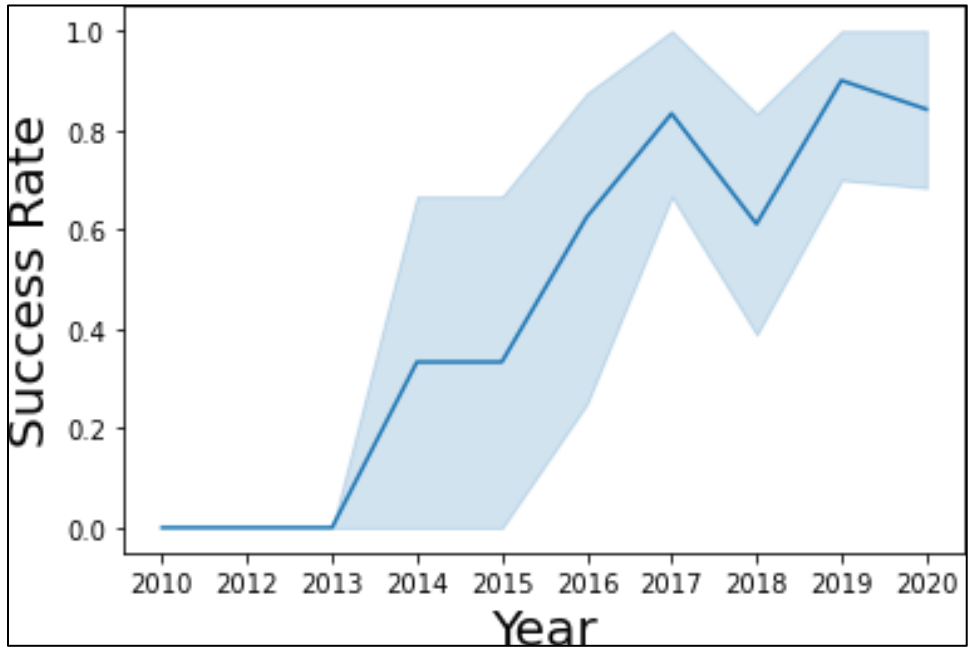
Payload vs. Orbit type



Green indicates successful launch; Purple indicates unsuccessful launch.

- Payload mass seems to correlate with orbit
- LEO and SSO seem to have relatively low payload mass
- The other most successful orbit VLEO only has payload mass values in the higher end of the range

Launch Success Yearly Trend



95% confidence interval (light blue shading)

Success rate increased since 2013 with a slight dip in 2018

Success in recent years stands at around 80%

EDA WITH SQL

EXPLORATORY DATA ANALYSIS WITH SQL DB2
INTEGRATED IN PYTHON WITH SQL ALCHEMY

All Launch Site Names

```
In [4]: %%sql
        SELECT UNIQUE LAUNCH_SITE
        FROM SPACEXDATASET;

         * ibm_db_sa://ftb12020:***@0c77d6f
        Done.
```

Out[4]:

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

- 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 Beginning with `CCA`

In [5]:
```
%%sql
SELECT *
FROM SPACEXDATASET
WHERE LAUNCH_SITE LIKE 'CCA%'
LIMIT 5;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.appdomain.cloud:31198/bludb
Done.
```

Out[5]:

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

First five entries in database with Launch Site name beginning with CCA.

Total Payload Mass from NASA

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS SUM_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE CUSTOMER = 'NASA (CRS)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
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

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD_MASS_KG
FROM SPACEXDATASET
WHERE booster_version = 'F9 v1.1'
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86
Done.
```

avg_payload_mass_kg
2928

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 Pad Landing Date

```
%%sql
SELECT MIN(DATE) AS FIRST_SUCCESS
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (ground pad)';
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81
Done.
```

first_success
2015-12-22

This query returns the first successful ground pad landing date.

First ground pad landing wasn't until the end of 2015.

Successful landings in general appear starting 2014.

Successful Drone Ship Landing with Payload Between 4000 and 6000

```
%%sql
SELECT booster_version
FROM SPACEXDATASET
WHERE landing__outcome = 'Success (drone ship)' AND payload_mass__kg_ BETWEEN 4001 AND 5999;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databas
Done.
```

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 exclusively.

Total Mission Outcome

```
%%sql
SELECT mission_outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
GROUP BY mission_outcome;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-
Done.
```

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

This query returns a count of each mission outcome.

SpaceX appears to achieve its mission outcome nearly 99% of the time.

Interestingly, one launch has an unclear payload status and unfortunately one failed in flight.

Boosters that Carried Maximum Payload

```
%%sql
SELECT booster_version, PAYLOAD_MASS__KG_
FROM SPACEXDATASET
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXDATASET);
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1
Done.
```

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

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 Failed Drone Ship Landing Records

```
%%sql
SELECT MONTHNAME(DATE) AS MONTH, landing__outcome, booster_version, PAYLOAD_MASS__KG_, launch_site
FROM SPACEXDATASET
WHERE landing__outcome = 'Failure (drone ship)' AND YEAR(DATE) = 2015;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg.databases.app
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.

There were two such occurrences.

Ranking Counts of Successful Landings Between 2010-06-04 and 2017-03-20

```
%%sql
SELECT landing__outcome, COUNT(*) AS no_outcome
FROM SPACEXDATASET
WHERE landing__outcome LIKE 'Succes%' AND DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY landing__outcome
ORDER BY no_outcome DESC;
```

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-86b520b87518.bs2io90l08kqb1od8lcg
Done.
```

landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

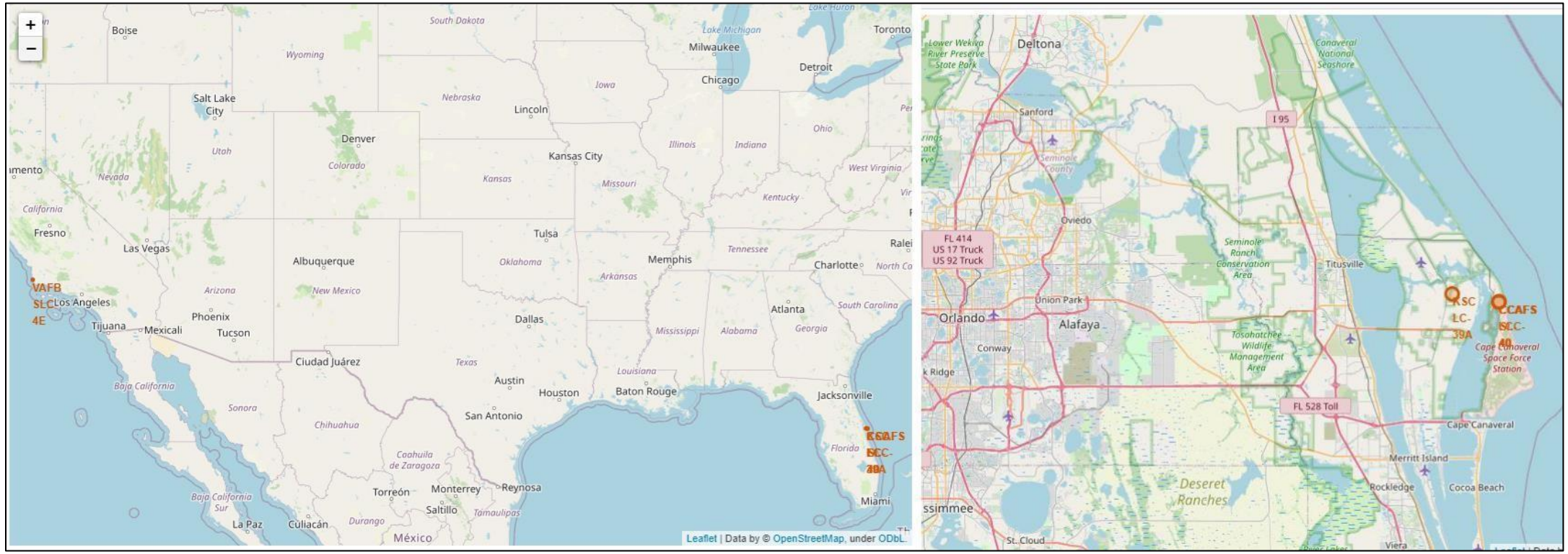
This query returns a list of successful landings and between 2010-06-04 and 2017-03-20 inclusively.

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

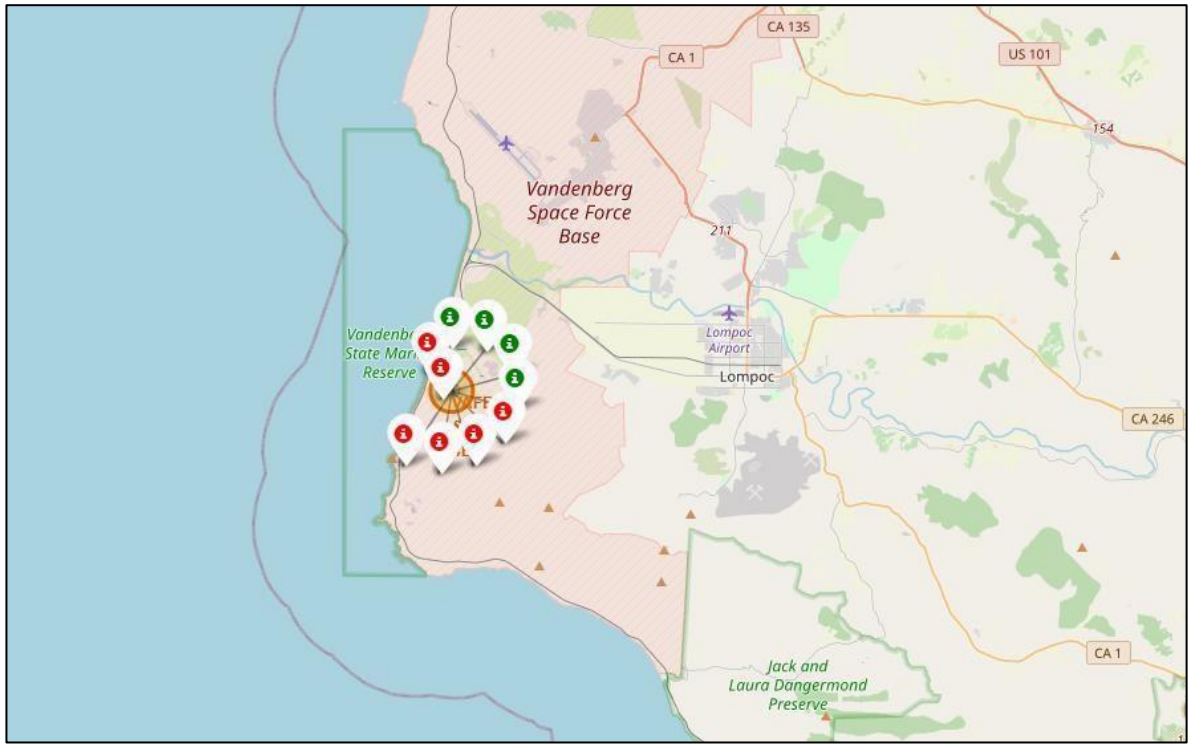
Interactive Map with Folium

Launch Site Locations



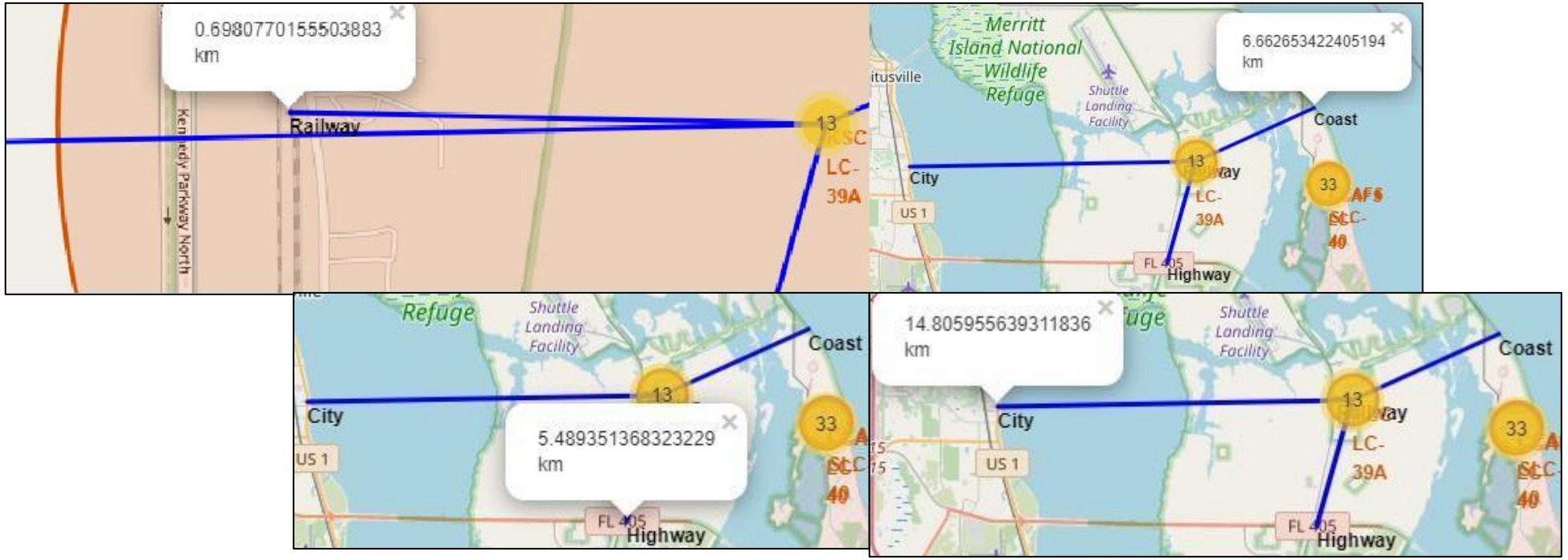
The left map shows all launch sites relative US map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

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.

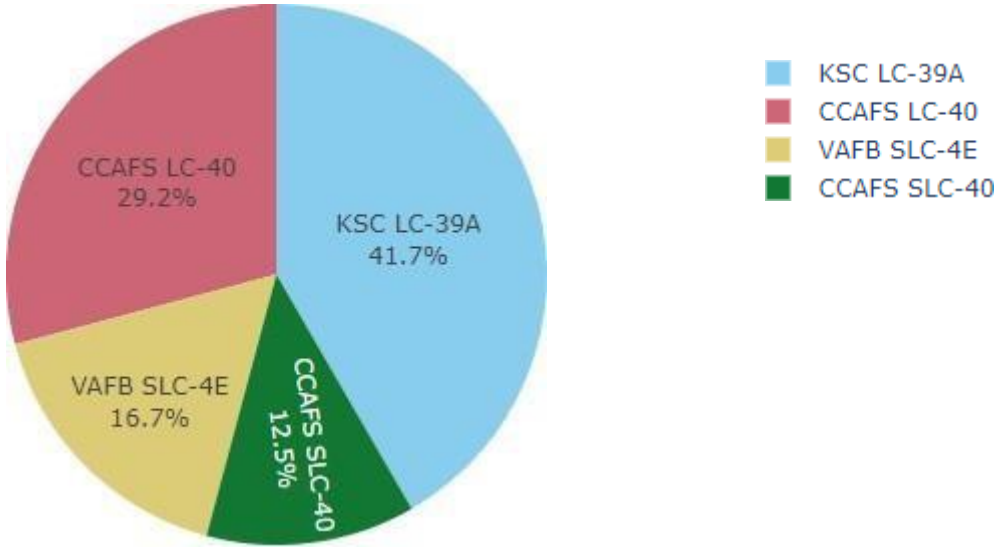
Key Location Proximities



Using KSC LC-39A as an example, launch sites are very close to railways for large parts transportation. Launch sites are close to highways for human and supply transport. 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.

Build a Dashboard with Plotly Dash

Successful Launches Across Launch Sites

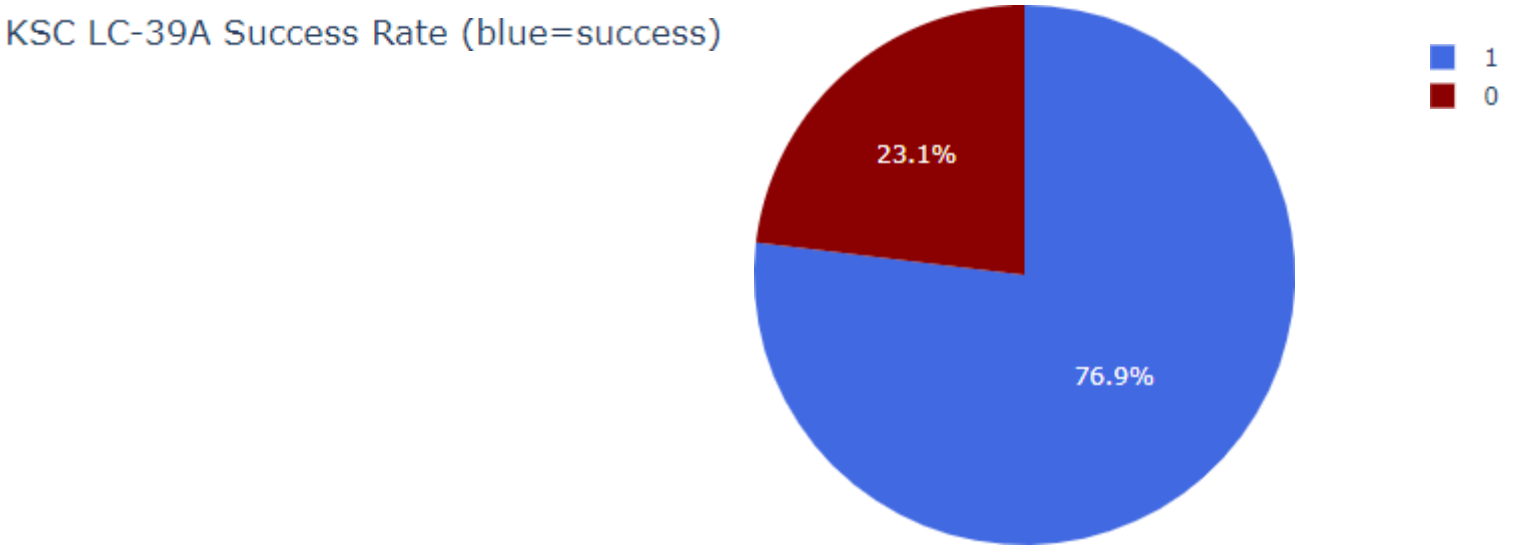


This is the distribution of successful landings across all launch sites.

CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same number of successful landings, but a majority of the successful landings were 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.

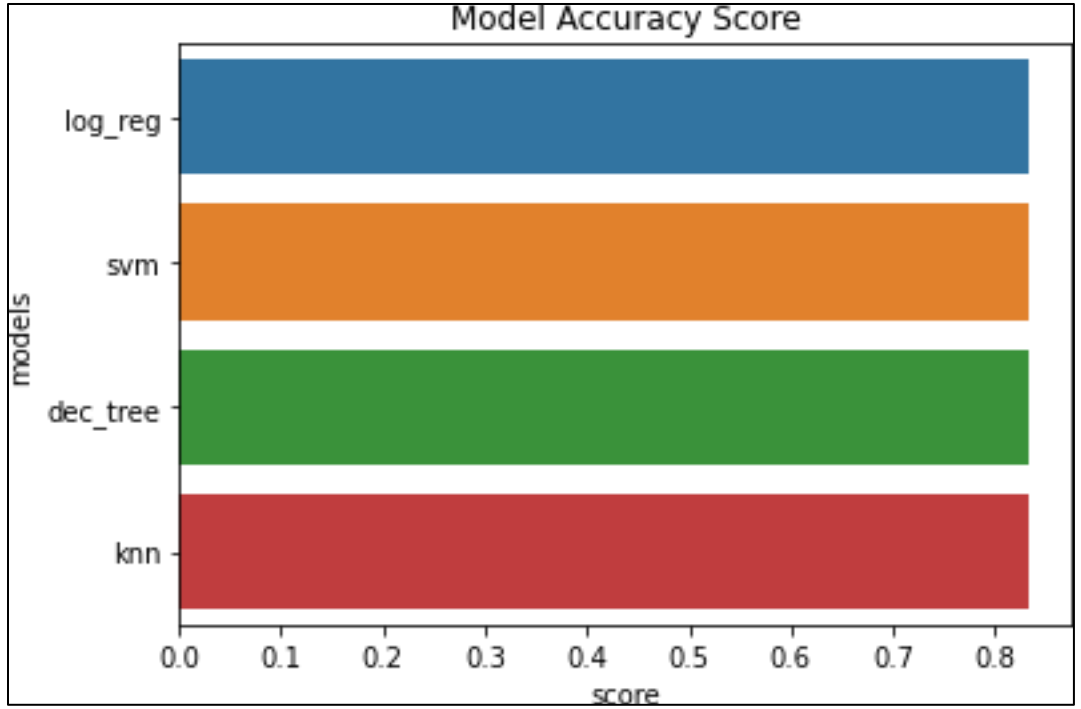
Payload Mass vs. Success vs. Booster Version Category

Plotly dashboard has a Payload range selector. However, this is set from 0-10000 instead of the max Payload of 15600.
Class indicates 1 for successful landing and 0 for failure.
Scatter plot also 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.

Predictive Analysis (Classification)

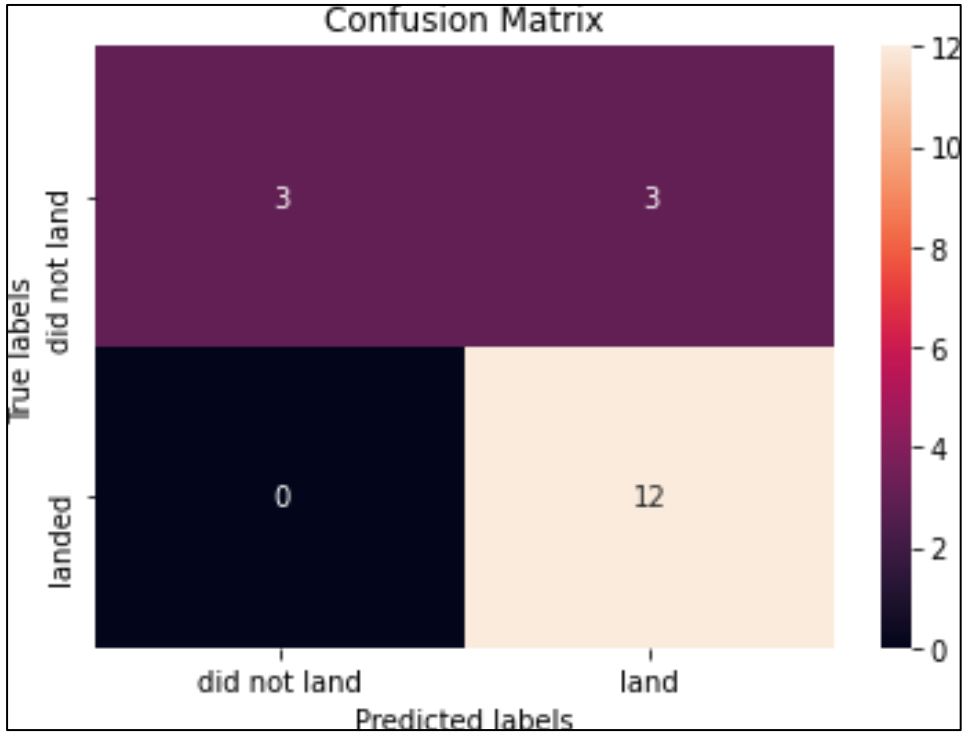
GRIDSEARCHCV(CV=10) ON LOGISTIC REGRESSION, SVM, DECISION TREE, AND KNN

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



Correct predictions are on a diagonal from top left to bottom right.

- Since all models performed the same for the test set, the confusion matrix is the same across all models. 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.

CONCLUSION

- ✓ Our task was to develop a machine learning model for SpaceY, aiming to compete with SpaceX.
- ✓ The objective of the model is to predict the likelihood of a successful Stage 1 landing, potentially saving around $100 million USD.
- ✓ We utilized data from a public SpaceX API and scraped information from the SpaceX Wikipedia page.
- ✓ After processing the data, we created labels and stored the information in a DB2 SQL database.
- ✓ We then developed a machine learning model, which achieved an accuracy of 83%.
- ✓ This allows the company to decide whether to proceed with the launch, potentially reducing risks and costs.
- ✓ To further enhance the model's performance, additional data collection would be beneficial to fine-tune the model and improve its accuracy.

APPENDIX

GitHub repository URL: https://github.com/Stephen-507/Python-

Special Thanks To:

Instructors

Coursera

Thank you