Data Science Capstone Project

JAVIER MOR

https://github.com/4r741

Summary I

1. Data Collection

- Collected data from the public SpaceX API and SpaceX Wikipedia page.

2. Data Labeling

- Created a 'class' column to classify successful landings.

3. Exploration

- Explored the data using SQL, visualizations, folium maps, and dashboards.

4. Feature Selection

- Gathered relevant columns to be used as features.

5. Data Transformation

- Changed all categorical variables to binary using one-hot encoding.

Summary II

6. Data Standardization

- Standardized data.

7. **Machine Learning Models**

- Utilized four machine learning models: Logistic Regression, Support Vector Machine,
- Decision Tree Classifier, and K Nearest Neighbors.
- Conducted GridSearchCV to find the best parameters for these models.

8. Model Evaluation

- Visualized accuracy scores of all models.

9. Results

- All models produced similar results with an accuracy rate of about 83.33%.
- Noted that all models tended to over-predict successful landings.

10. Recommendations

- Suggested that more data is needed for better model determination and improved accuracy.

Background

Entering the Commercial Space Age, SpaceX has revolutionized the space industry with competitive pricing, offering a significant advantage over competitors, like Space Y ($62 million vs. $165 million USD). SpaceX's cost-effectiveness is attributed to its capability to recover a part of the rocket, particularly Stage 1.

Challenge

Space Y, aiming to compete with SpaceX, has commissioned us to develop a machine learning model. The primary objective is to predict the successful recovery of Stage 1, a crucial factor in cost optimization and competitiveness in the commercial space launch market.

Approach

1.Data Collection
Gathered relevant data from SpaceX's public API and Wikipedia, including information on launches, landing outcomes, and other pertinent features.

2.Data Labeling
Created a 'success' label to indicate the successful recovery of Stage 1.

3.Exploratory Data Analysis (EDA)
Conducted in-depth EDA to understand patterns, trends, and potential features influencing successful recoveries.

4.Feature Selection
Identified key features contributing to the prediction of successful Stage 1 recovery.

5.Machine Learning Models
Developed and trained machine learning models, including Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors.

6.Model Evaluation
Utilized performance metrics, visualizations, and accuracy scores to evaluate and compare the effectiveness of the models.

Recommendations

Provided insights and recommendations based on model results.
Emphasized the need for continuous data collection and model refinement to enhance predictive accuracy.

Impact

The successful implementation of the machine learning model will empower Space Y with the ability to make informed decisions, optimize costs, and enhance the likelihood of successful Stage 1 recoveries, ultimately strengthening its competitive position in the commercial space launch industry.

Data Collection Methodology I

Data Collection

Integration of information from the SpaceX public API and the SpaceX Wikipedia page.

Data collection and organisation process.

Data cleaning

Application of data cleaning techniques to ensure consistency and quality.

Landing Classification

Creation of labels to classify landings as successful or unsuccessful.

Exploratory Data Analysis (EDA)

Use of visualisations and SQL queries to explore patterns and trends in the data.

Data Collection Methodology II

Interactive Visual Analytics

Implementation of tools such as Folium and Plotly Dash for interactive visual analytics.

Predictive Analytics

Application of classification models to predict the success of landings.

Model Fitting

Optimisation of models through GridSearchCV to improve their performance.

Impact

This comprehensive methodology allows not only to collect meaningful data from a variety of sources, but also to explore, visualise and predict key events related to SpaceX landings.

The interactive and predictive approach provides a solid foundation for informed decision making and continuous process optimisation in the space industry.

Methodology

Overview Of Data Collection
Wrangling
Visualization
Dashboard
Model Methods

DATA COLLECTION

Data Collection Process

Data collection was carried out using a combination of requests to the SpaceX public API and web scraping techniques from the SpaceX Wikipedia entry.

SpaceX API Data

Significant data was obtained from the API, including columns such as FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, and Latitude.

Web Scraping data from Wikipedia

Data collected via web scraping from the SpaceX Wikipedia entry table included columns such as Flight No., Launch site, Payload, PayloadMass, Orbit, Customer, Launch outcome, Version Booster, Booster landing, Date and Time.

Flow of the process

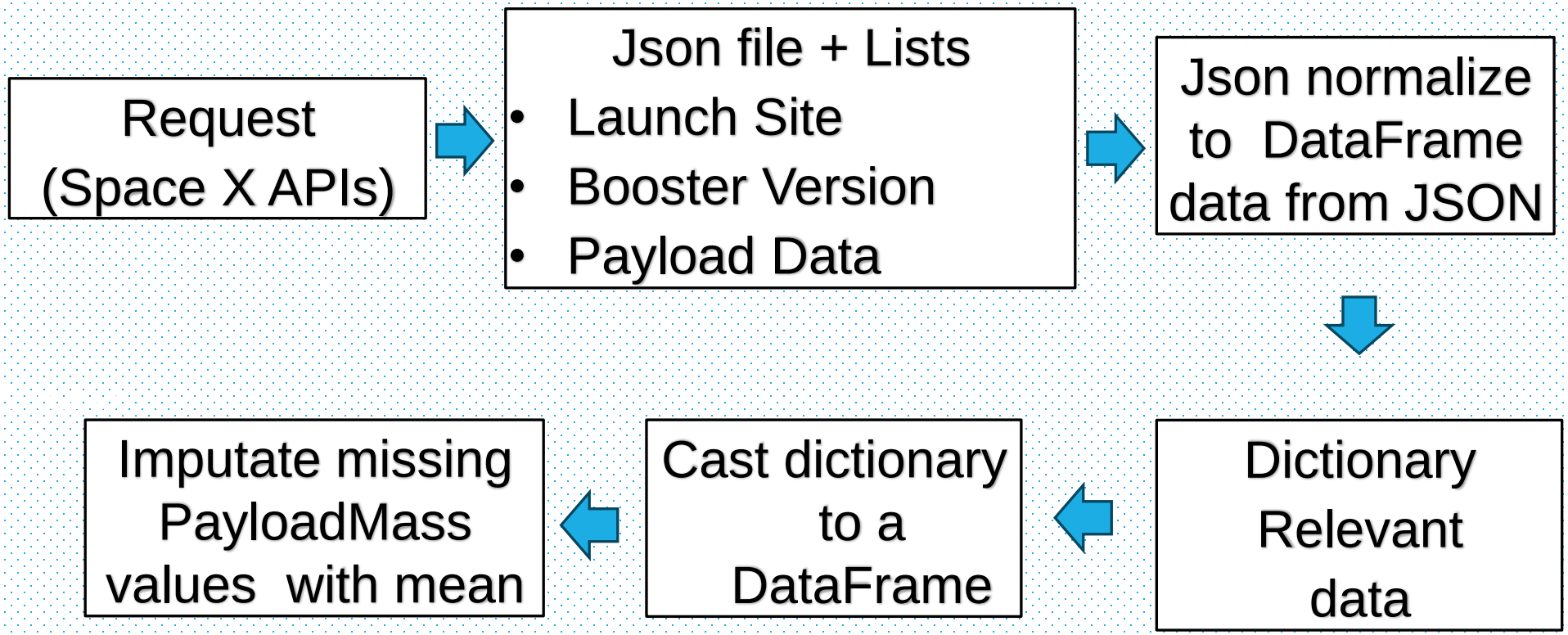
The following slide will show the flowchart of the data collection process from the API.

The subsequent slide will present the flowchart of the web scraping process.

This comprehensive approach ensures relevant and comprehensive data is obtained from multiple sources for further analysis and application in the project.

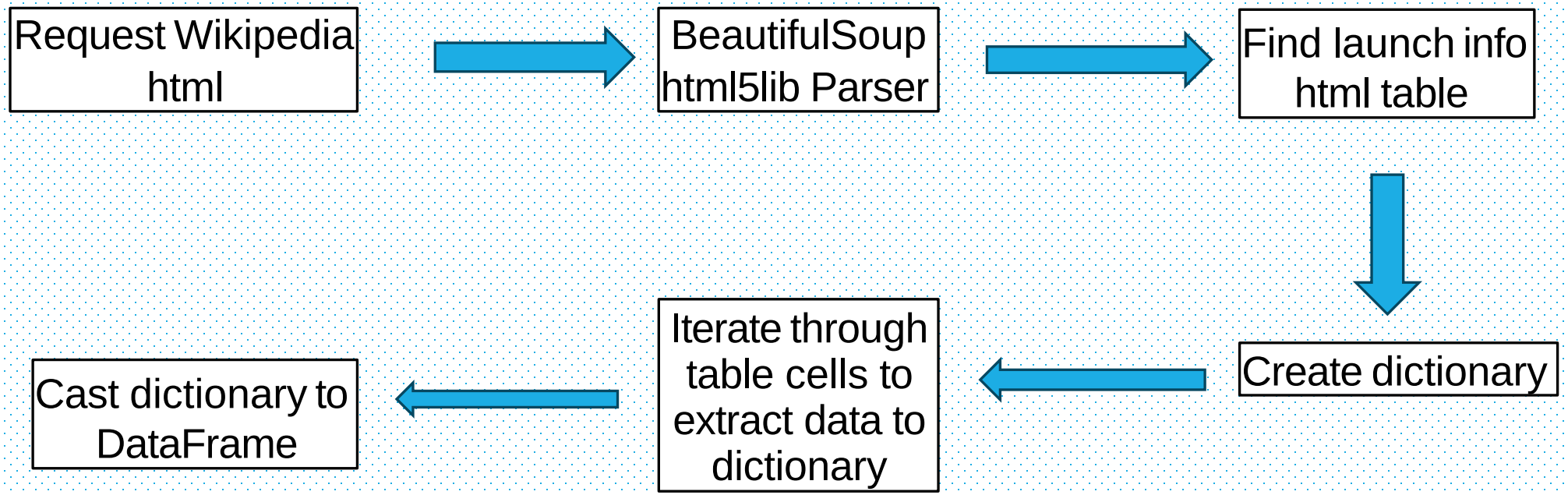
Data Collection

SpaceX API



Data Collection

Web Scraping



Data Wrangling

A training tag has been created for landing results, where:

- ✓ Successful landing is labeled as 1.
- ✓ Failed landing (failure) is labeled as 0.

The 'Outcome' column consists of two components: 'Mission Outcome' and 'Landing Location'.

A new column of training labels called 'class' has been created with the following values:

- ✓ If 'Mission Outcome' is True and 'Landing Location' is ASDS, RTLS or Ocean, it is set to 1.
- ✓ For all other combinations (None None, False ASDS, None ASDS, False Ocean, False RTLS), it is set to 0.

This approach provides a clear, binary training label for classification modeling, making it easier to train machine learning models to predict the outcome of the SpaceX landings.

EDA With Data Visualization I

An Exploratory Data Analysis was carried out on the variables Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.

Graphics Used:

1. Flight Number vs. Payload Mass
2. Flight Number vs. Launch Site
3. Payload Mass vs. Launch Site
4. Orbit vs. Success Rate
5. Flight Number vs. Orbit
6. Payload vs. Orbit
7. Annual Success Trend

EDA With Data Visualization II

Types of Graphics Used

Scatter plots, line graphs, and bar plots were used to compare the relationships between variables. The objective was to determine the existence of relationships that could be used in training the machine learning model.

Results

- Patterns and key relationships between variables were identified.
- The trend of success over the years was evaluated.
- Effective visualization provides valuable information for building the predictive model.

This comprehensive analysis lays the foundation for deep understanding of the data and its application in machine learning model training.

EDA With SQL I

Data Integration and SQL Queries

The dataset was loaded into an IBM DB2 database, and queries were performed using SQL and Python integration. The queries were performed with the goal of gaining a deeper understanding of the dataset.

Types of Queries Performed

1. Launch site name information.
2. Mission Results.
3. Customer payload sizes and thruster versions.
4. Landing results.

EDA With SQL II

Results and Benefits

- Improved Understanding of the Dataset: Queries provided detailed information on several key aspects of the dataset.
- Facilitated Exploratory Analysis: Integration with a database allowed for more complex analysis and targeted information.
- Relevant Information for Modeling: Query results can be relevant for building and improving predictive models.

This approach to data integration and querying enhances analytical capabilities and lays the foundation for informed decisions in the modeling and prediction process.

Build An Interactive Map With Folium I

Folium Maps and Location Visualization

Folium maps were used to mark launch sites, successful and unsuccessful landings, and an example of proximity to key locations: rail, road, coast and city.

Objectives of Map Use

1. **Launch Site Locations**
   Folium maps help visualize the geographic distribution of launch sites.

2. **Successful and Failed Landings**
   Marking successful and failed landings provides a geospatial view of the results.

3. **Proximity to Key Locations**
   Explore proximity to important features such as railroads, highways, coastline and cities.

Build An Interactive Map With Folium II

Benefits of Visualization

- **Understanding of Key Location Locations**
  Enables understanding of why launch sites are located where they are.

- **Visualization of Successful Landings**
  Provides a visual representation of successful landings in relation to their location.

This spatial approach to visualization adds a significant layer of understanding to data analysis and can be key to strategic decision making for future launches.

Build A Dashboard With Plotly Dash I

Dashboard Implementation

A dashboard has been implemented that includes a pie chart and a scatter plot.

Control Panel Functionalities:

1. **Pie Chart**
    i. Allows selection of the distribution of successful landings at all launch sites.
    ii. Provides the option to display individual success rates for each launch site.
    iii. Effectively visualizes the success rate at the launch sites.

Build A Dashboard With Plotly Dash II

2. **Scatter Plot**

   i. Select from all sites or an individual site.

   ii. Uses a slider to adjust the payload mass between 0 and 10000 kg.

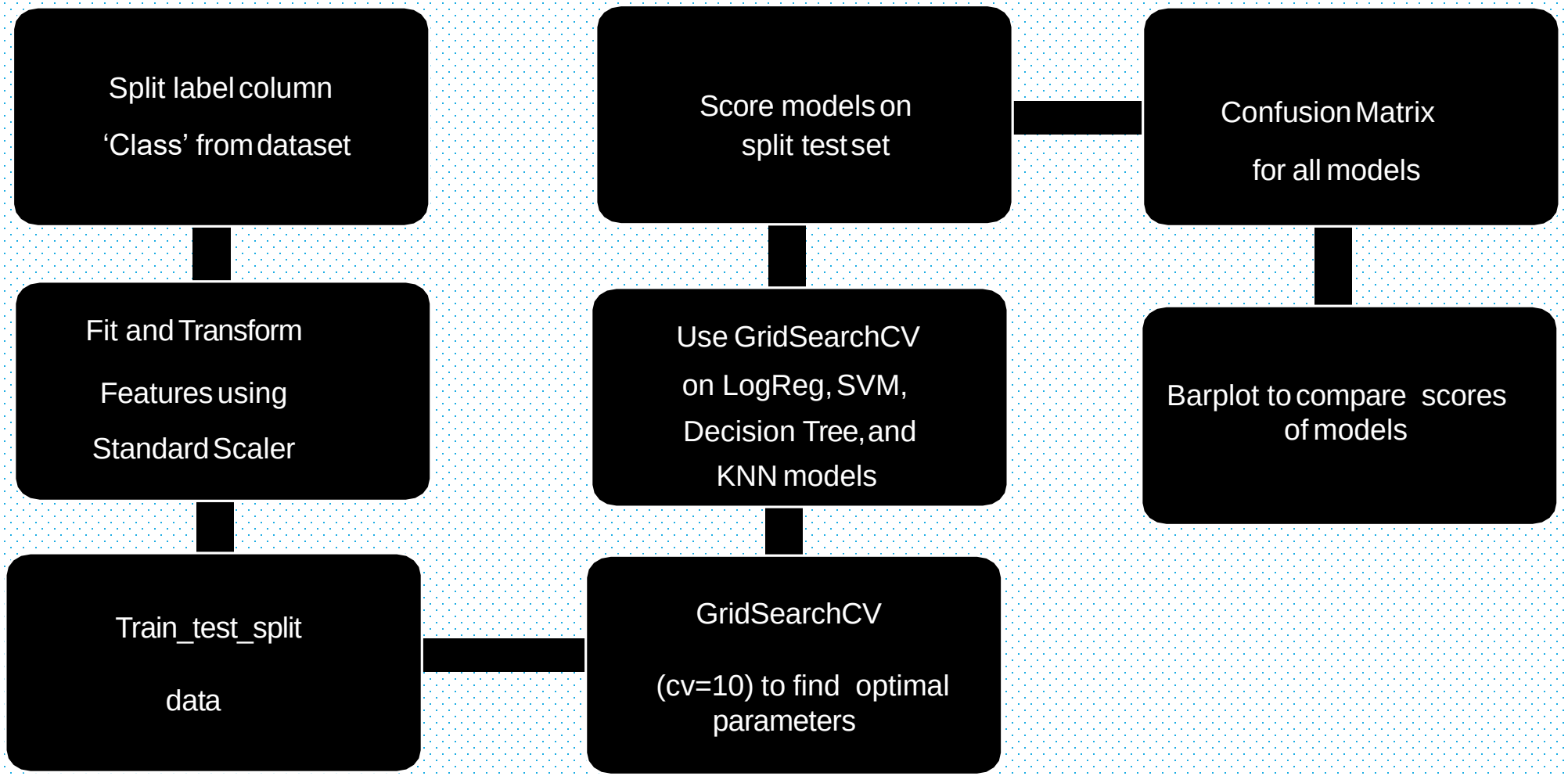
   iii. Allows visualization of how success varies as a function of launch sites, payload mass, and propellant version category.

Purpose of the Dashboard

The dashboard provides an interactive tool to explore and understand the variability in launch success as a function of several key factors, such as launch site, payload mass, and propellant version. This facilitates informed decision making and the identification of relevant patterns in the data collected.

This focus on interactive visualization enhances the understanding of the data and its application in strategic decision making for Space Y.

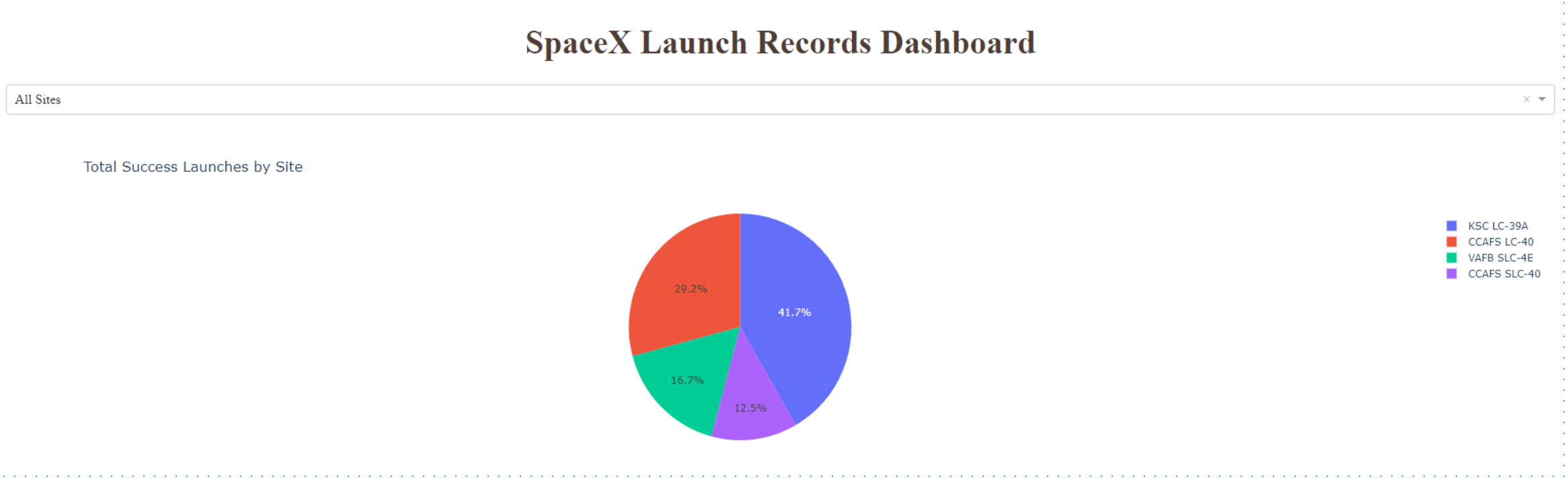
Predictive Analysis



Results

Dashboard Preview

The provided preview showcases the Plotly dashboard, offering a glimpse into various sections that unveil the outcomes of Exploratory Data Analysis (EDA) through visualization, EDA utilizing SQL, an Interactive Map powered by Folium, and concluding with the model's results boasting an approximate accuracy of 83%.



Results

EDA with Visualization:

Visual outcomes from data exploration, spotlighting significant patterns and relationships.

EDA with SQL

Exploratory analysis leveraging SQL queries to extract specific insights from the database.

Interactive Map with Folium

An interactive geospatial visualization depicting relevant locations.

Model Results

Assessment and presentation of the machine learning model's outcomes, achieving an 83% accuracy.

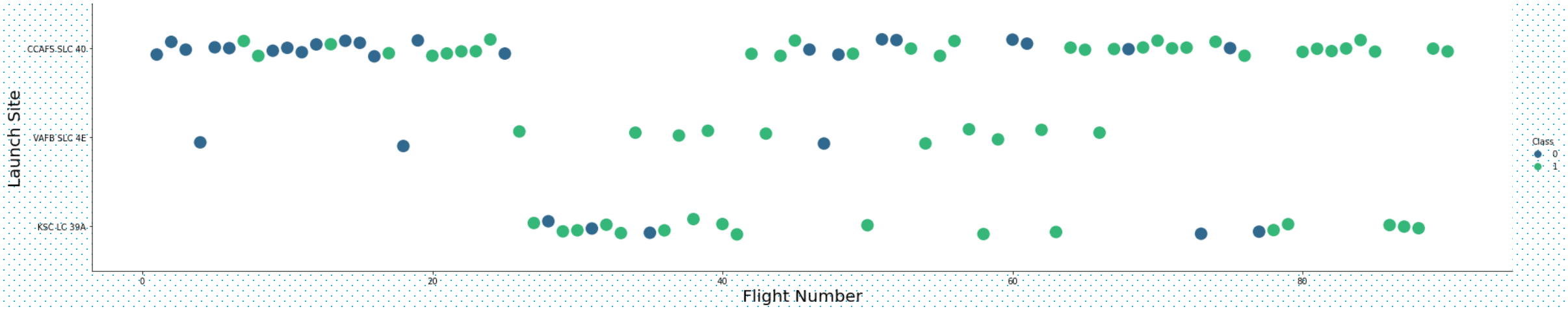
Dashboard Significance:

- Offers a comprehensive visual representation of both data and model outcomes.
- Facilitates effective interpretation and communication of findings to stakeholders.
- Enables swift decision-making based on the information presented.
- This dashboard reflects the culmination of the undertaken work, spanning from initial analysis to the implementation of the predictive model.

EDA with Visualization

Exploratory Data Analysis With Seaborn Plots

Flight Number Vs. Launch Site



Green indicates successful launch **Purple indicates unsuccessful launch.**

Success Rate Trends Over Time

The graphical representation, particularly in the context of Flight Number, indicates a noteworthy surge in success rates over time. Notably, there seems to be a significant breakthrough around the 20th flight, contributing substantially to the overall success rate improvement.

Primary Launch Site

The data analysis points towards Cape Canaveral Air Force Station (CCAFS) emerging as the predominant launch site. This inference is drawn from the observation that CCAFS exhibits the highest volume of launches compared to other sites.

Flight Number Vs. Launch Site II

Implications

Temporal Success Enhancement

The discerned breakthrough around the 20th flight suggests a potential turning point or improvement in the launch procedures, contributing to heightened success rates.

Strategic Significance of CCAFS

The prominence of CCAFS as the primary launch site highlights its strategic importance in the overall operations, potentially serving as a focal point for future endeavors.

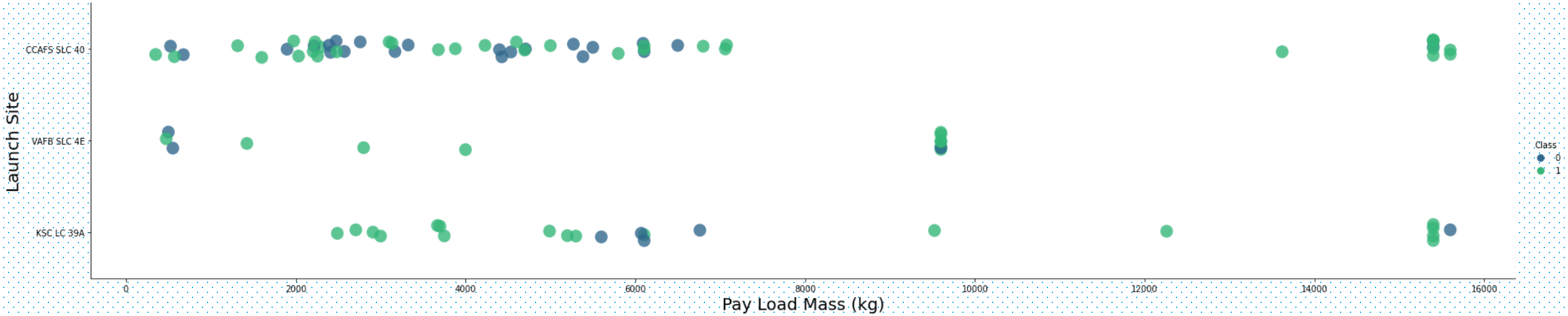
Actionable Insights

- Consider conducting an in-depth analysis of the specific circumstances surrounding the 20th flight to identify contributing factors to the success surge.

- Further explore and optimize procedures at CCAFS, leveraging its significance as the primary launch site.

These insights, derived from visual analysis, lay the groundwork for targeted investigations and strategic decision-making.

Payload vs. Launch Site



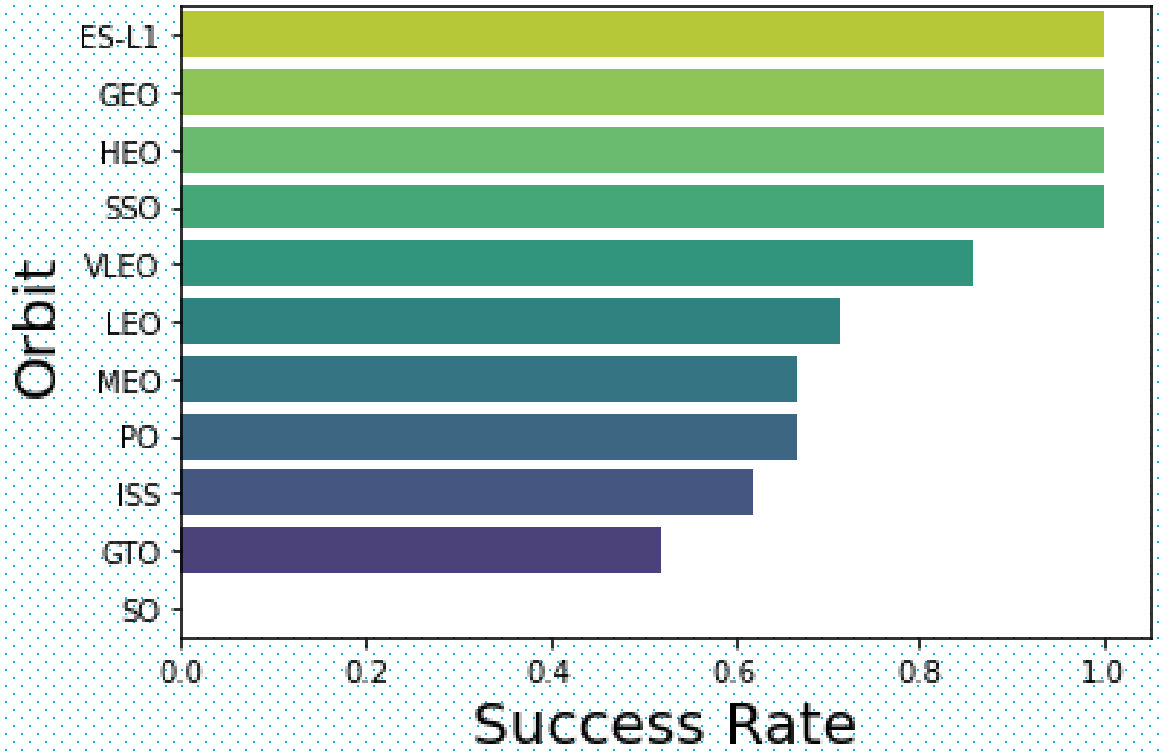
Green indicates successful launch

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 with
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)

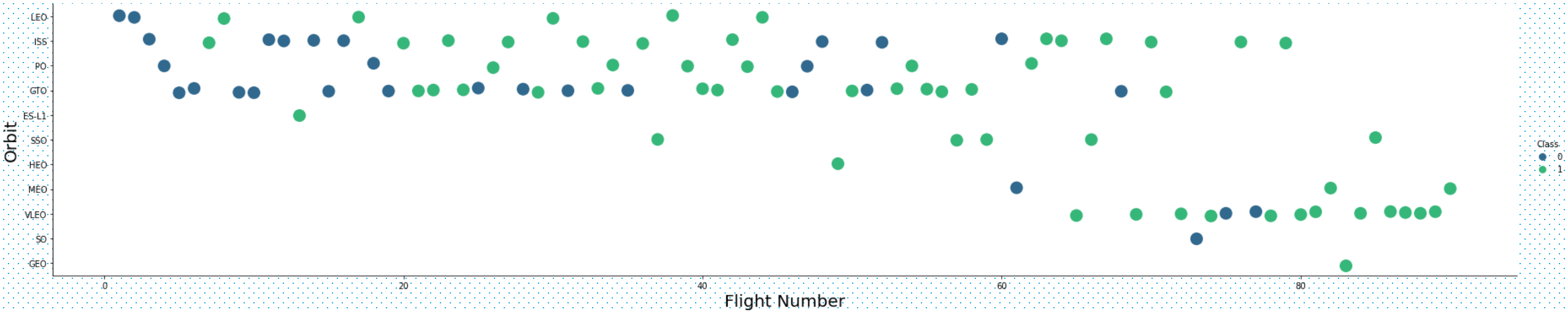
SSO (5) has 100% success rate

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. ORBITTYPE



Green indicates successful launch; Purple indicates unsuccessful launch.

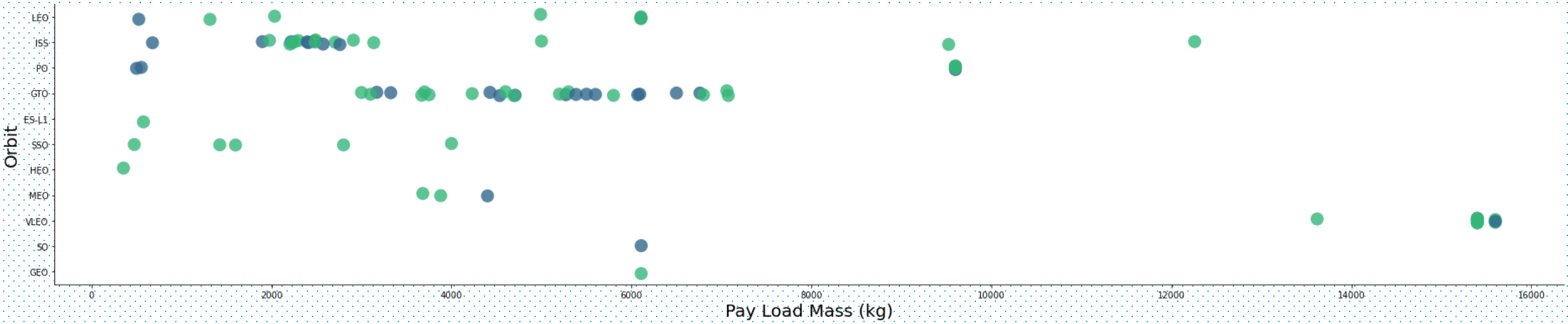
Launch Orbit preferences changed over Flight Number.

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



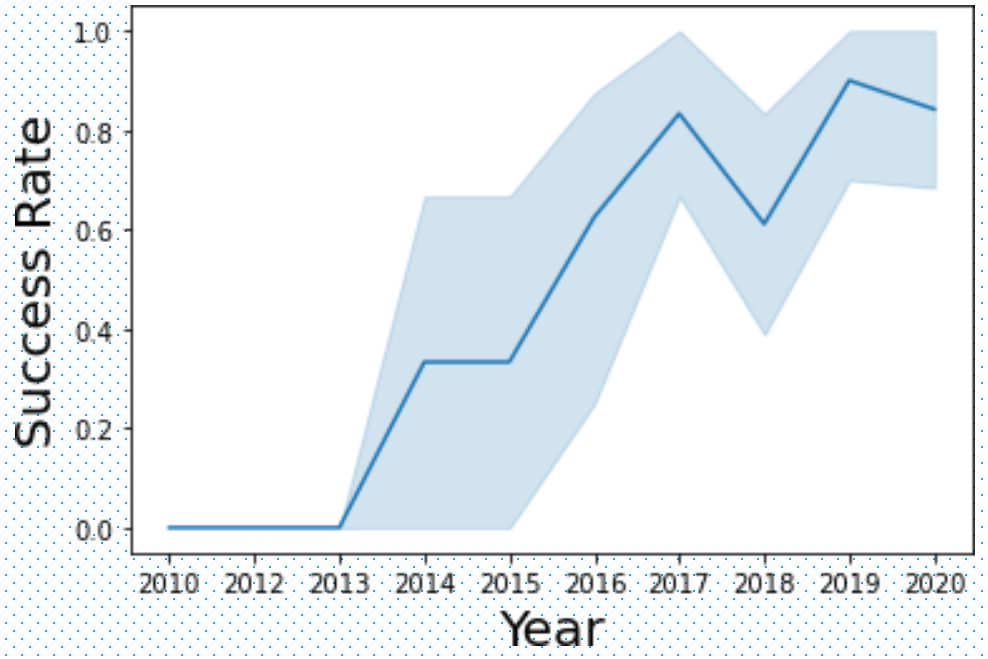
Green indicates successful launch; Purple indicates unsuccessful launch.

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 generally increases over time since 2013 with a slight dip in 2018

Success in recent years at around 80%

EDA with SQL

EXPLORATORY DATA ANALYSIS WITH SQL DB2

INTEGRATED IN PYTHON WITH SQLALCHEMY

ALL LAUNCH SITE NAMES

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

```
 * ibm_db_sa://ftb12020:***@0c77d6f2
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 F9V1.1

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

```
 * ibm_db_sa://ftb12020:***@0c77d6f2-5da9-48a9-81f8-8(
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.database
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 noninclusively.

TOTAL NUMBER OF EACH 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.

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 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 SUCCESSFULLANDINGS BETWEEN 2010-06-04 AND2017-03-20

```
%%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.
```

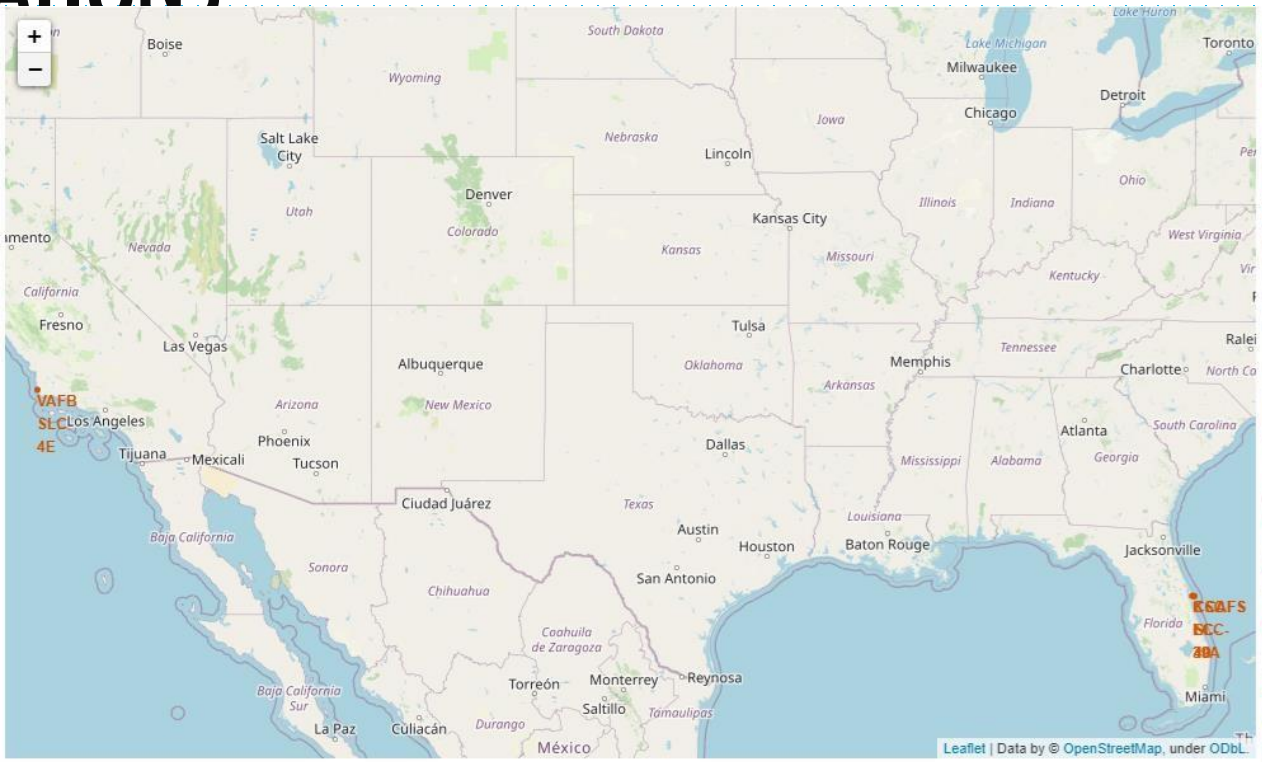
landing__outcome	no_outcome
Success (drone ship)	5
Success (ground pad)	3

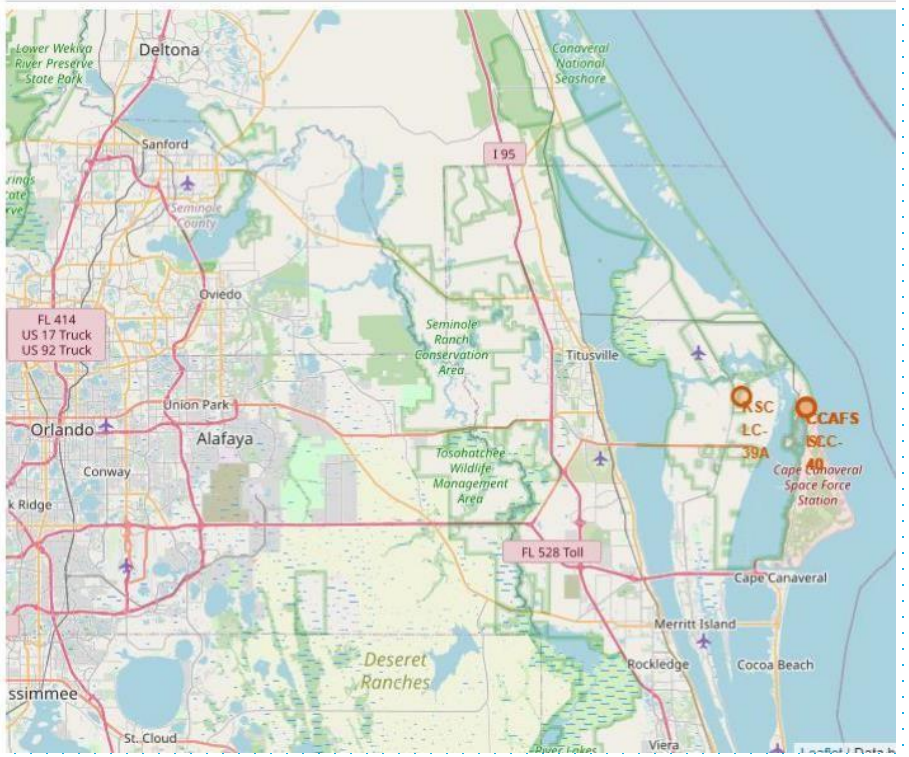
This query returns a list of successful landings and between 2010-06-04 and 2017-03-20 inclusively.

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

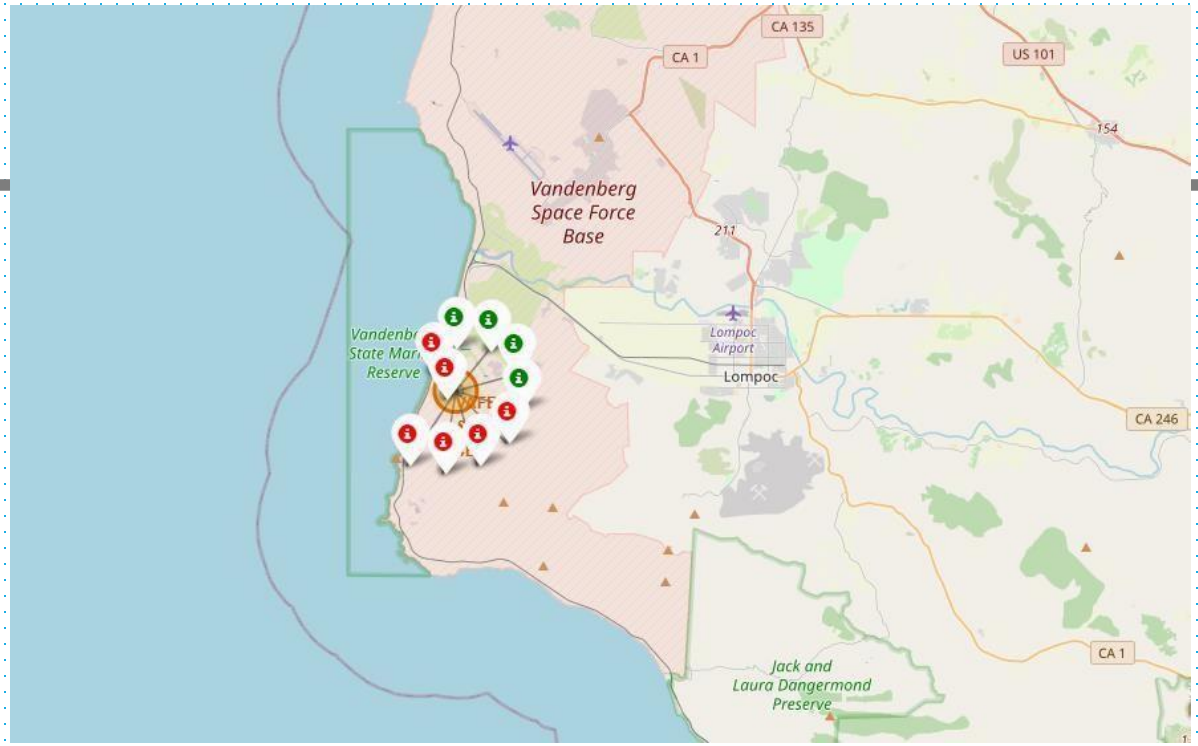
LAUNCH SITE LOCATIONS





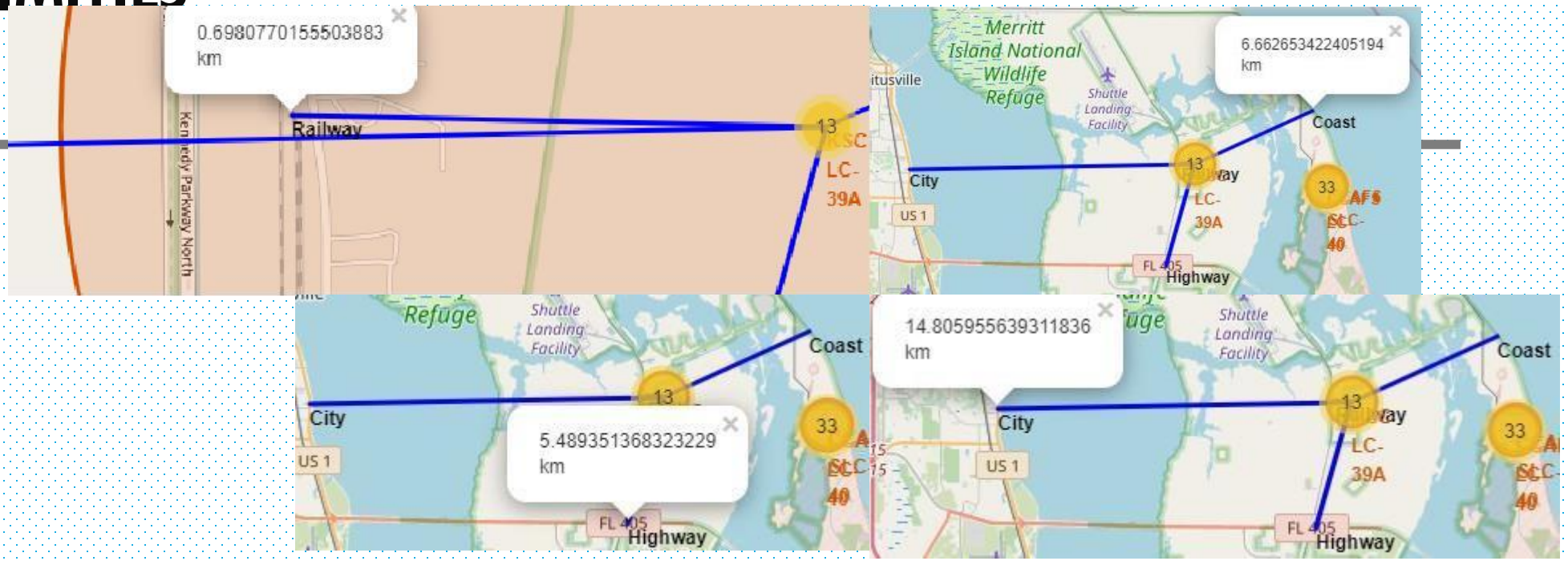
The left map shows all launch sites relative US map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

COLOR-CODED LAUNCH MARKERS



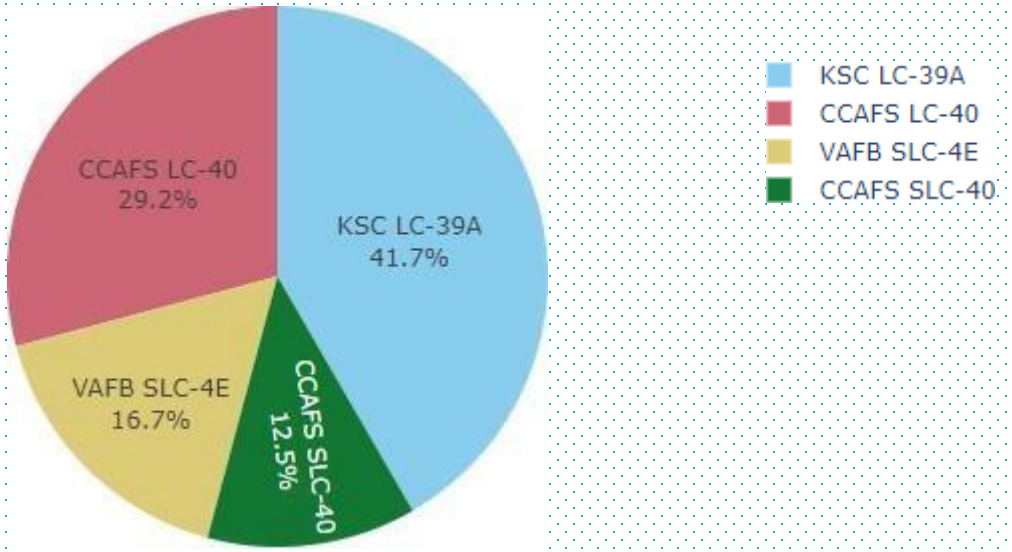
Clusters on Folium map can be clicked on to display each successful landing (green icon) and failed landing (red icon). In this example VAFB SLC-4E shows 4 successful landings and 6 failed landings.

KEY LOCATION PROXIMITIES



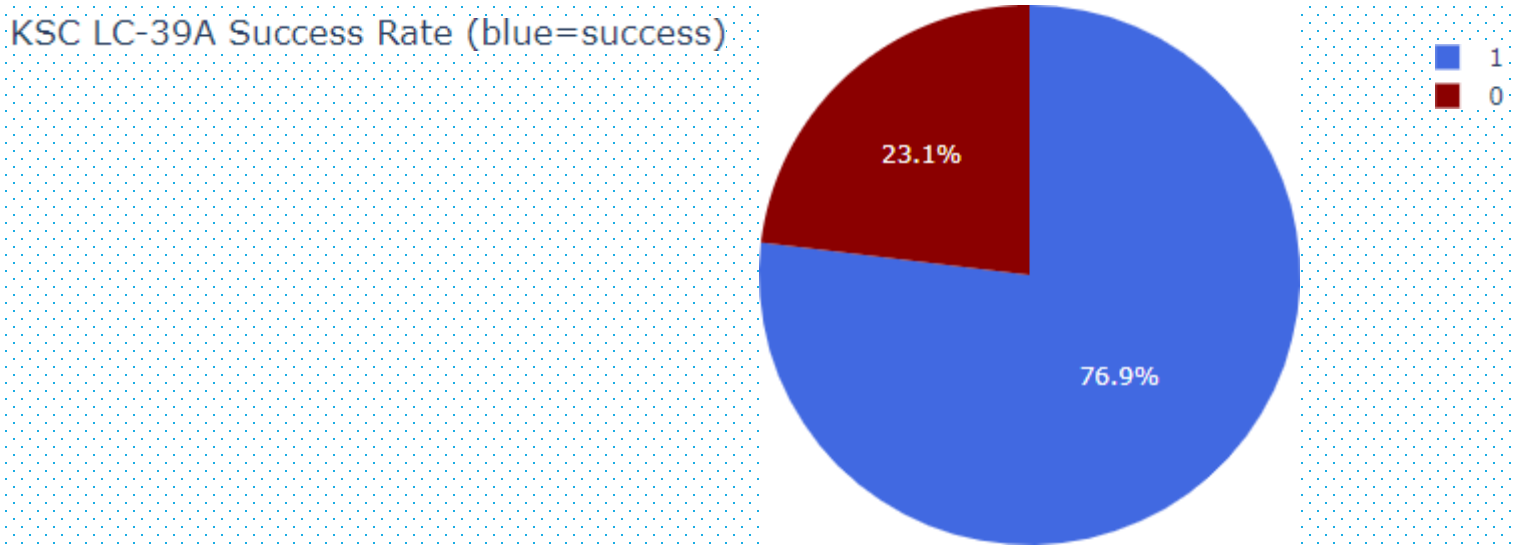
Using KSC LC-39A as an example, launch sites are very close to railways for large part and supply transportation. Launch sites are close to highways for human and supply transport. Launch sites are also close to coasts and relatively far from cities so that launch failures can land in the sea to avoid rockets falling on densely populated areas.

SUCCESSFUL LAUNCHES ACROSS LAUNCH SITES



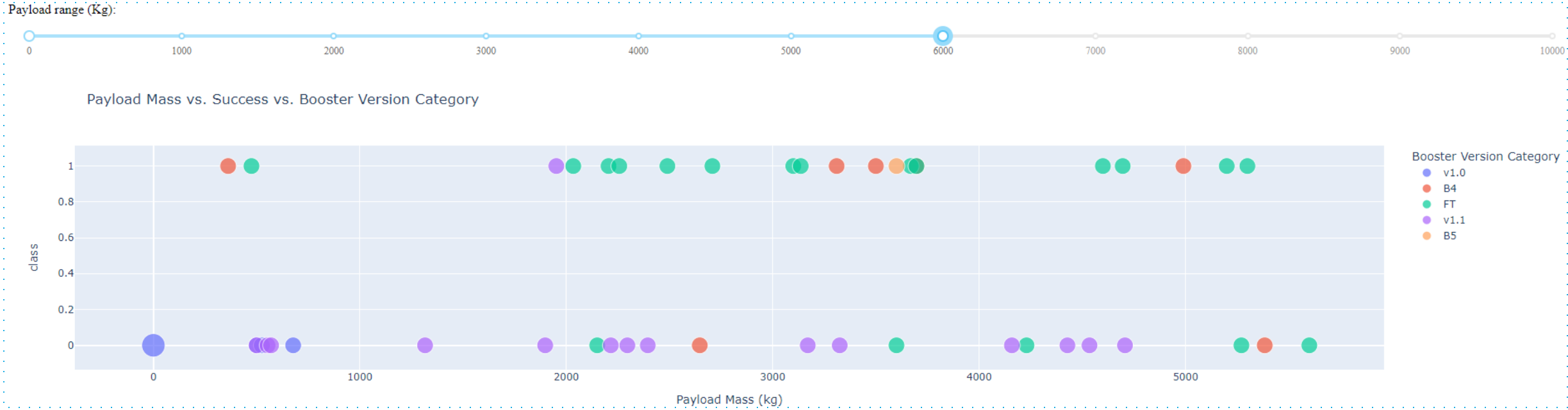
This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings where performed before the name change. VAFB has the smallest share of successful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

HIGHEST SUCCESS RATE LAUNCH SITE



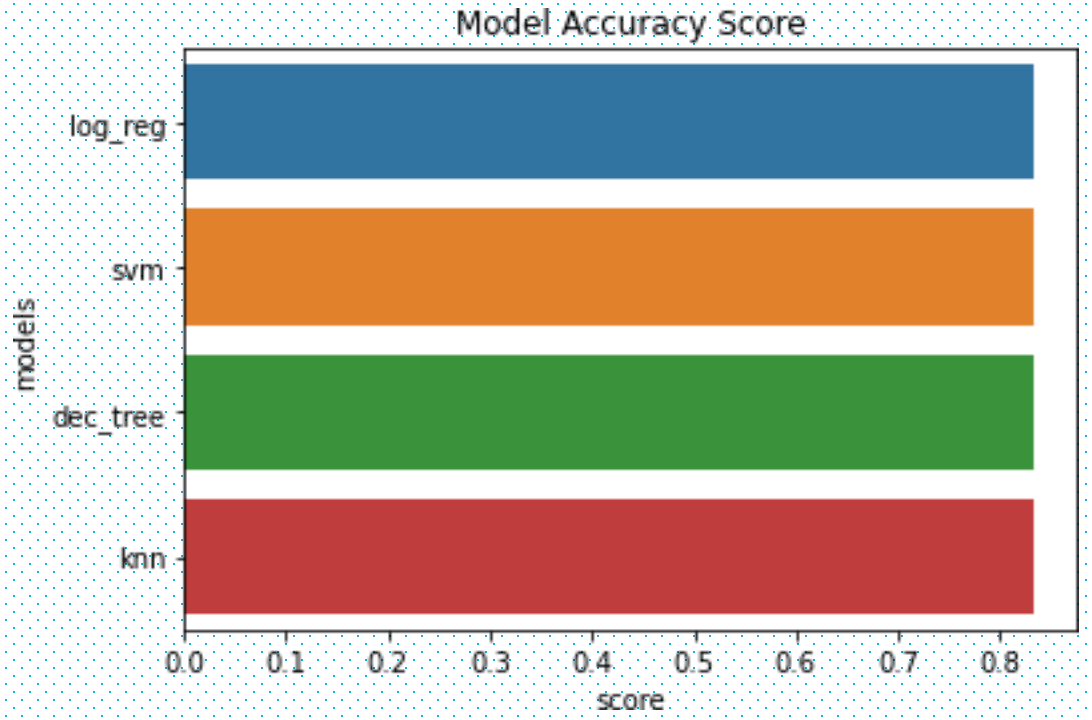
KSC LC-39A has the highest success rate with 10 successful landings and 3 failed landings.

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.

CLASSIFICATION ACCURACY



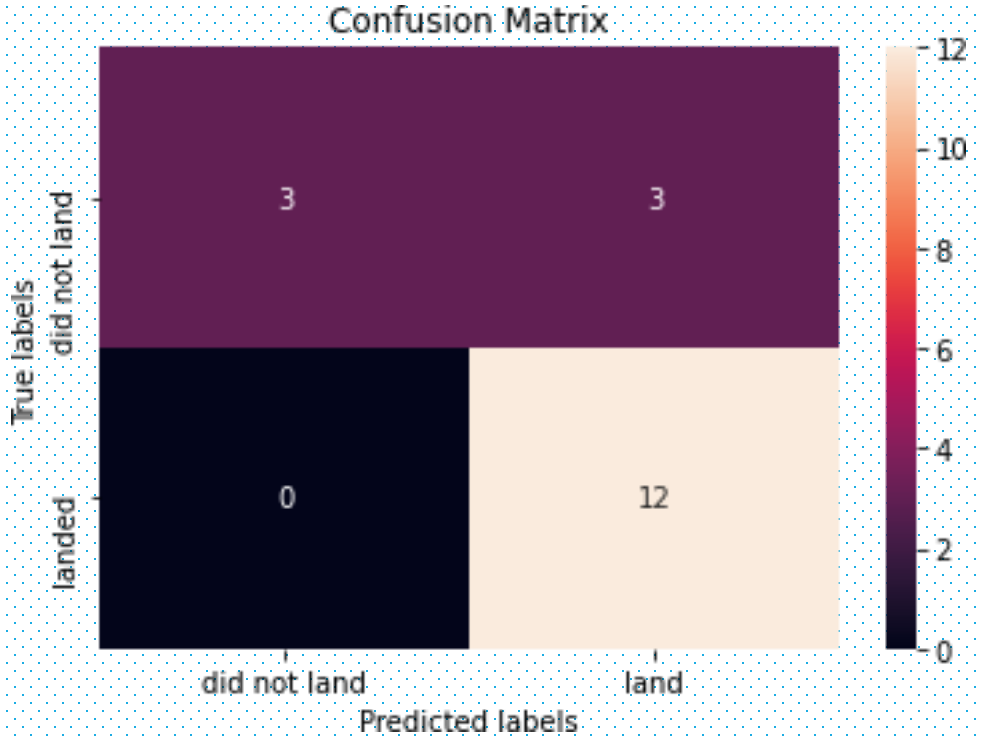
All models had virtually the same accuracy on the test set at 83.33% accuracy.

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.