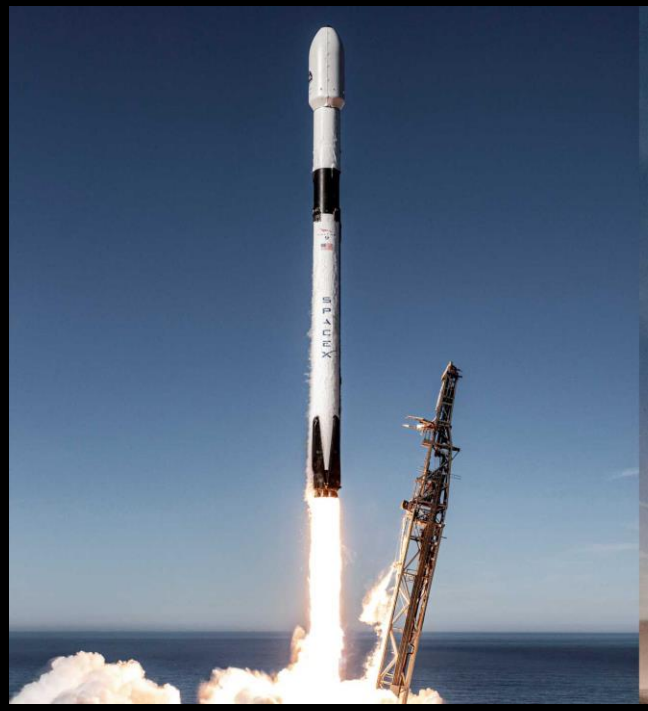
Space Y

The New Competitor of the Space Race

Nadir Tony Shawish

Outline



Executive Summary

- Data was collected from the public SpaceX API and SpaceX Wikipedia page.
- Created labels column ‘class’ classifies successful landings.
- Data explored using SQL
- Explored data was then visualized with folium maps and dashboards.
- Relevant columns were then gathered as features.
- Changed all categorical variables to binary using one hot encoding.
- Standardized the data
- Used GridSearchCV to find best parameters for machine learning models.

- Four machine learning models were produced: Logistic Regression, Support Vector Machine, Decision Tree Classifier, and K Nearest Neighbors. All produced similar results with accuracy rate of about 83.33%. All models over predicted successful landings.
- More data is needed for better model determination and accuracy.

Introduction

SpaceX Falcon 9 Rocket – The Verge

Background:

- Commercial Space Age has begun!
- SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars,
- other providers cost upward of 165 million dollars each,
- SpaceX can reuse the first stage (Stage 1) of the rocket

Problem:

- Space Y would like for us to develop a prediction model to predict successful Stage 1 recoveries

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

Methodology

-Data Collection & Wrangling,
-Data Visualization & Dashboard,
-Predictive Analysis

Data Collection Overview

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.

The next slide will show the flowchart of data collection from API and the one after will show the flowchart of data collection from webscraping.

Space X API Data Columns:

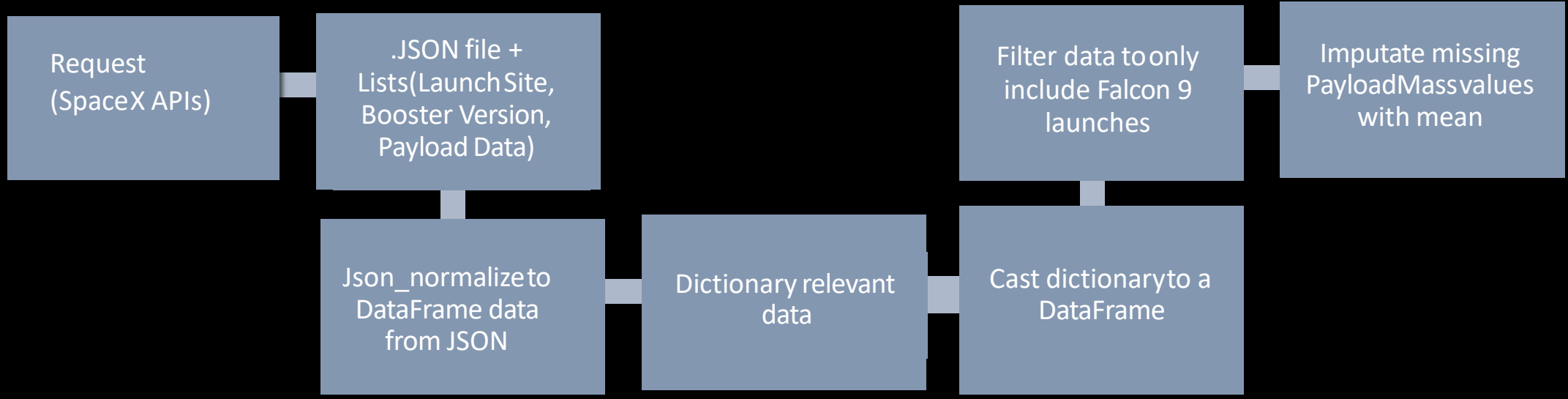
FlightNumber, Date, BoosterVersion, PayloadMass, Orbit, LaunchSite, Outcome, Flights, GridFins, Reused, Legs, LandingPad, Block, ReusedCount, Serial, Longitude, Latitude

Wikipedia Webscrape Data Columns:

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

Data Collection

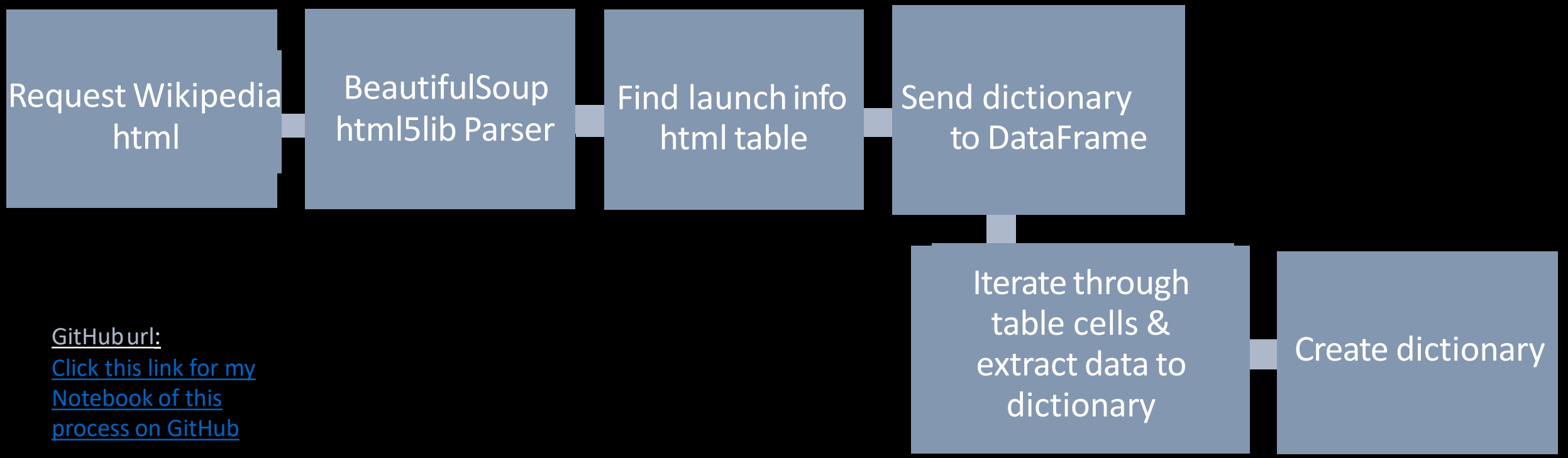
–SpaceX API Process:



GitHub url:
Click this link for my Notebook of this process on GitHub

Data Collection

Web Scraping
(through Wikipedia HTML)



GitHub url:
Click this link for my Notebook of this process on GitHub

Data Wrangling

- Create a training label with landing outcomes (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, 0 if otherwise.
- Value Mapping: True ASDS, True RTLS, & True Ocean – set to -> 1
- None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0

- GitHub url:Click this link for my Notebook of this process on GitHub

Exploratory Data Analysis with Data Visualization

EDA performed on variables:

-Flight Number, Payload Mass, Launch Site, Orbit, Class and Year.

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

-Visualization: Scatter plots, line charts, and bar plots were used to compare relationships between variables to decide if a relationship exists

-If so, they could be used in training the machine learning model

GitHub url:Click this link for my Notebook of this process on GitHub

Exploratory Data Analysis with SQL

-Loaded data set into IBM DB2 Database.

-Queried using SQL Python integration. (to get a better understanding of the dataset.)

-Queried info: launch site names, mission outcomes, pay load sizes of customers, booster versions, and landing outcomes

GitHub url:Click this link for my Notebook of this process on GitHub

Interactive Map
Folium

-Folium maps mark Launch Sites, successful/unsuccessful landings
-also mark 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 a location.

GitHub url:Click this link for my Notebook of this process on GitHub

Dashboard with Plotly Dash

-Dashboard includes a pie chart and a scatter plot.

-Pie chart: shows distribution of successful landings across all launch sites, can be selected to show individual launch site success rates.

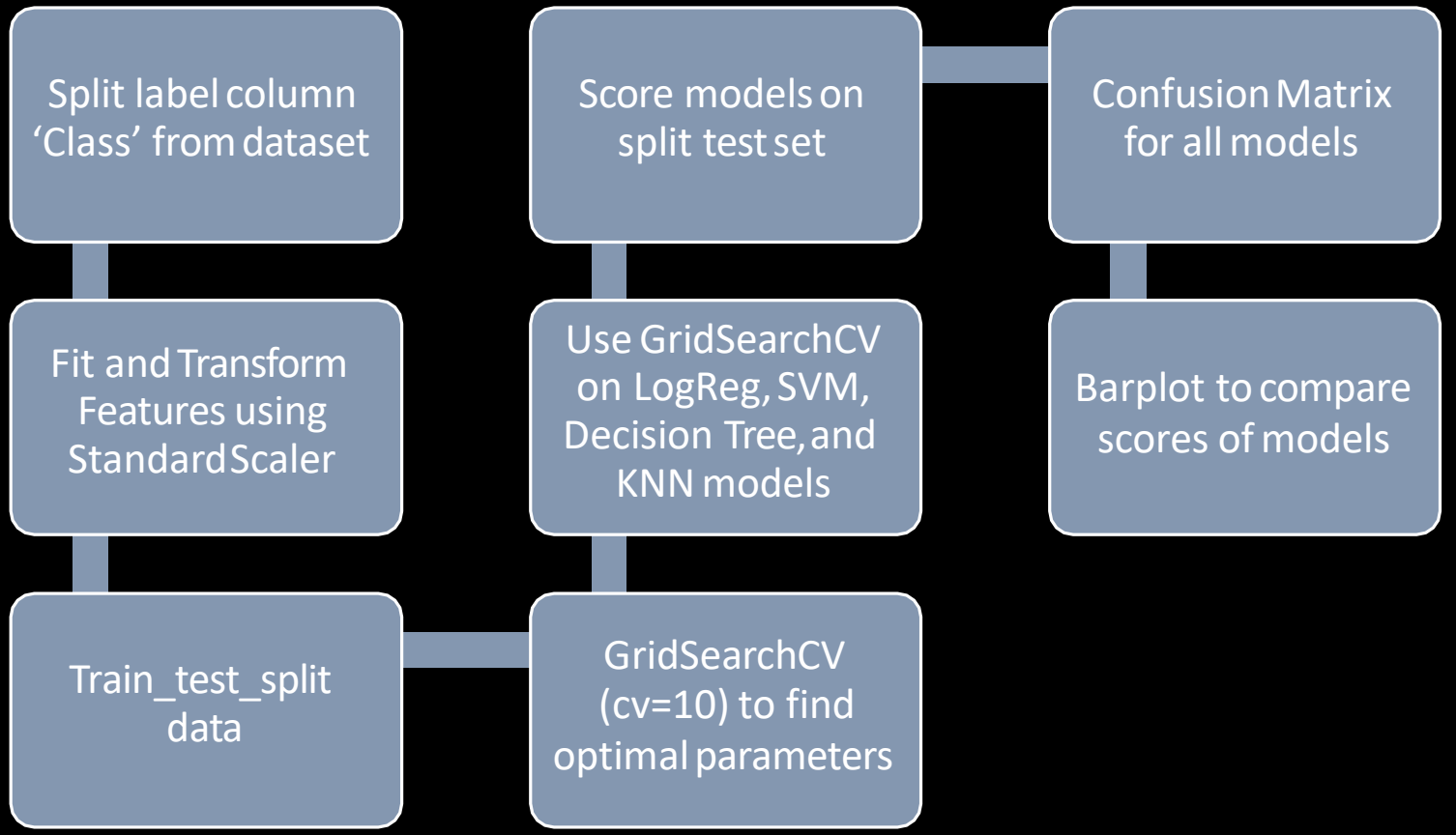
-Scatter plot: takes two inputs: All sites or individual site and payload mass on a slider between 0 and 10000 kg.

- This will help us see how success varies across launch sites, payload mass, and booster version category.

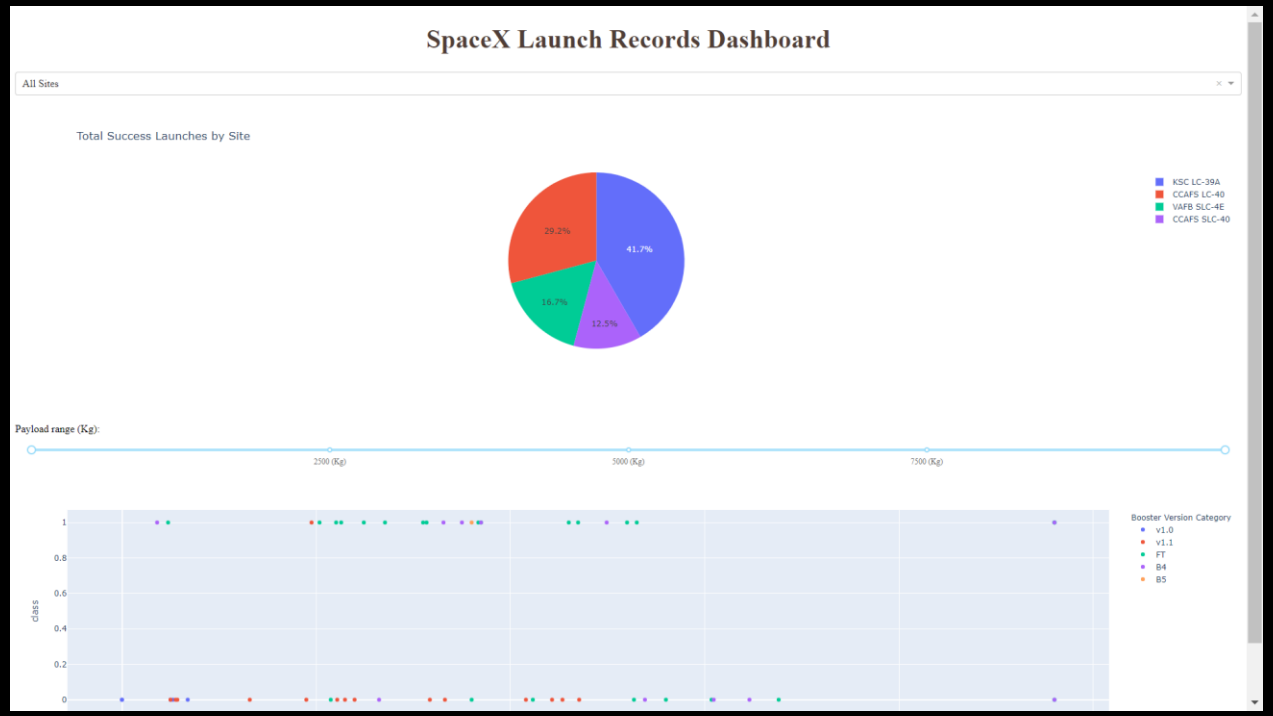
GitHub url:click this link to view python script of this process on GitHub

Predictive analysis (Classification)

GitHub url:click this link to view my Notebook of this process on GitHub



Results

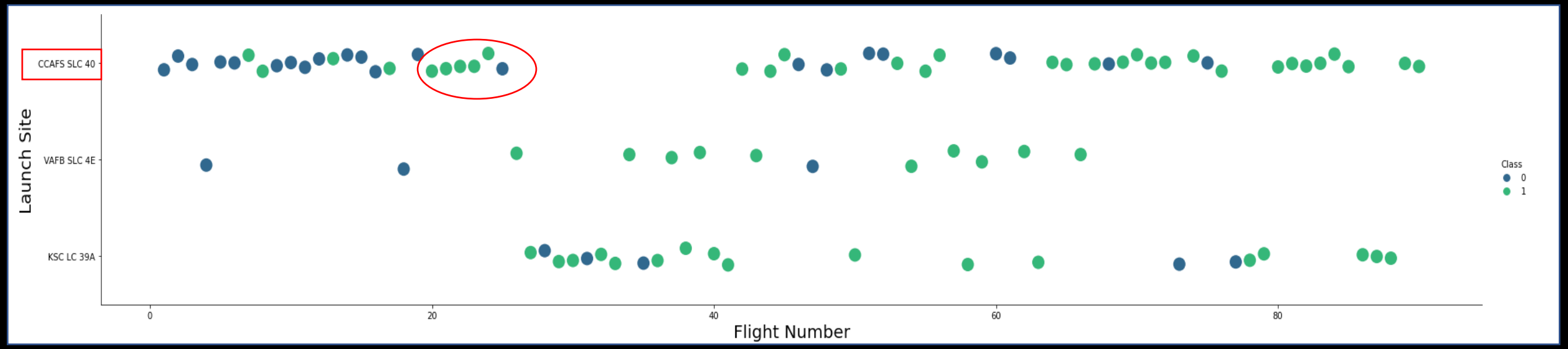


This is a preview of the Plotly dashboard. The following slides will show the results of EDA with visualization, EDA with SQL, Interactive Map with Folium, and the results of our model with about 83% accuracy.

EDA with Visualization

EXPLORATORY DATA ANALYSIS WITH SEABORN PLOTS

Flight Number vs. LaunchSite



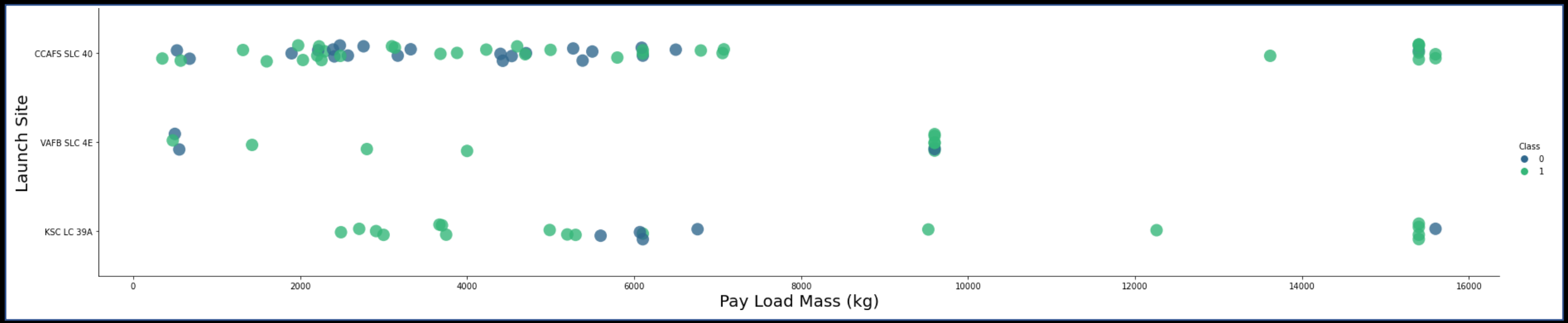
Green = successful launch, Blue = unsuccessful launch.

-Graphic suggests an increase in success rate over time (indicated in Flight Number).

-Likely a big breakthrough around flight 20 as there is a significantly increased success rate.

-CCAFS SLC 40 appears to be the main launch site as it has the most volume.

Payload vs. Launch Site



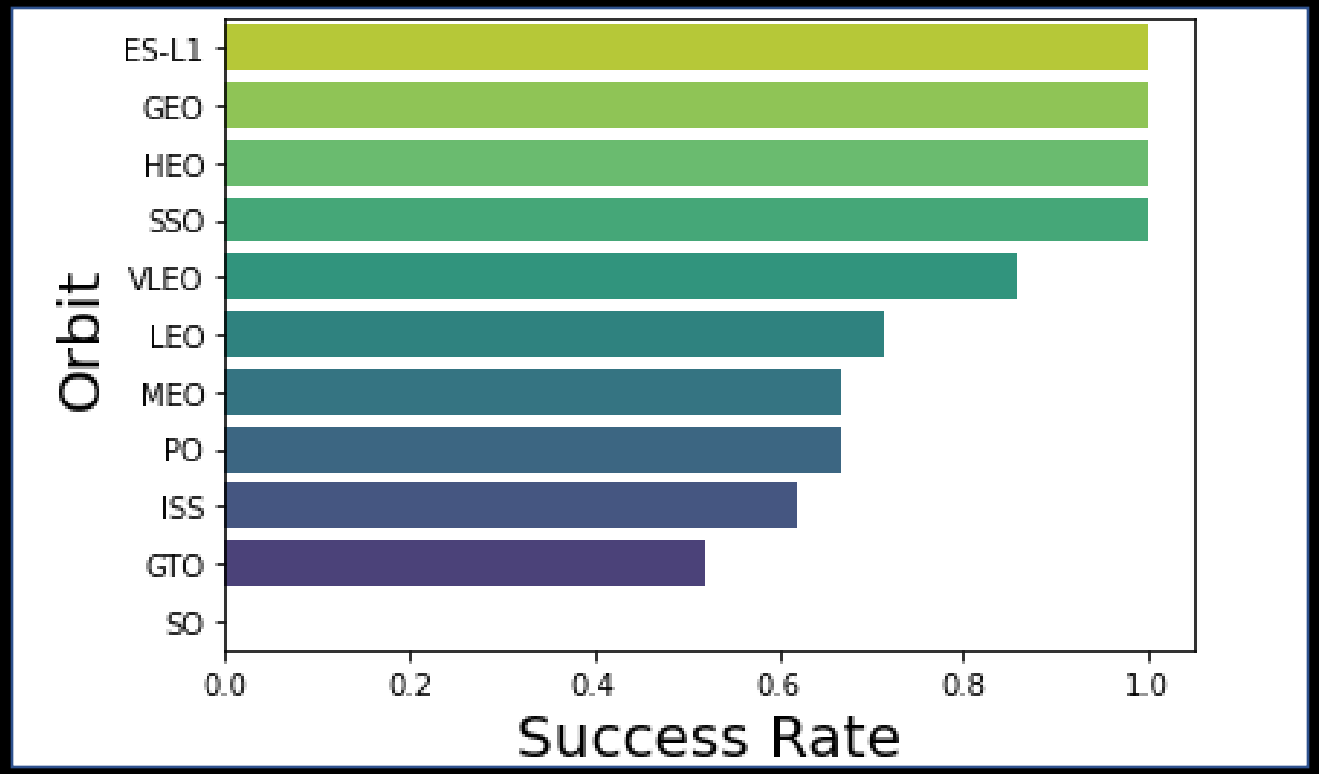
Green = successful launch, Blue = unsuccessful launch.

-Payload mass appears to fall mostly between 0-6000 kg.

-VAFB SLC 4E Launch Site does not have any Launches with a Pay Load Mass greater than 10000…

-CCAFS SLC 40 has more successful Launches of greater payloads.

Success rate vs. Orbit Type



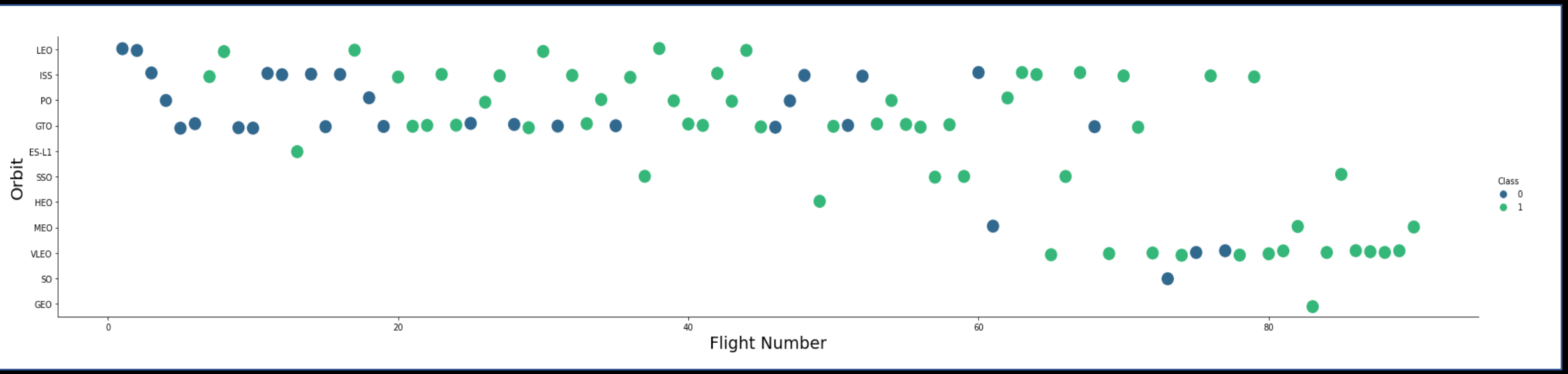
Success Rate Scale
decimal : percentage

-ES-L1, GEO, HEO, SSO have 100% success rates (sample sizes in parenthesis) 100% success rate

-SO has 0% success rate

-GTO has the around 50% success rate, but the largest sample

Flight Number vs. Orbit Type



Green = successful launch, Blue = unsuccessful launch.

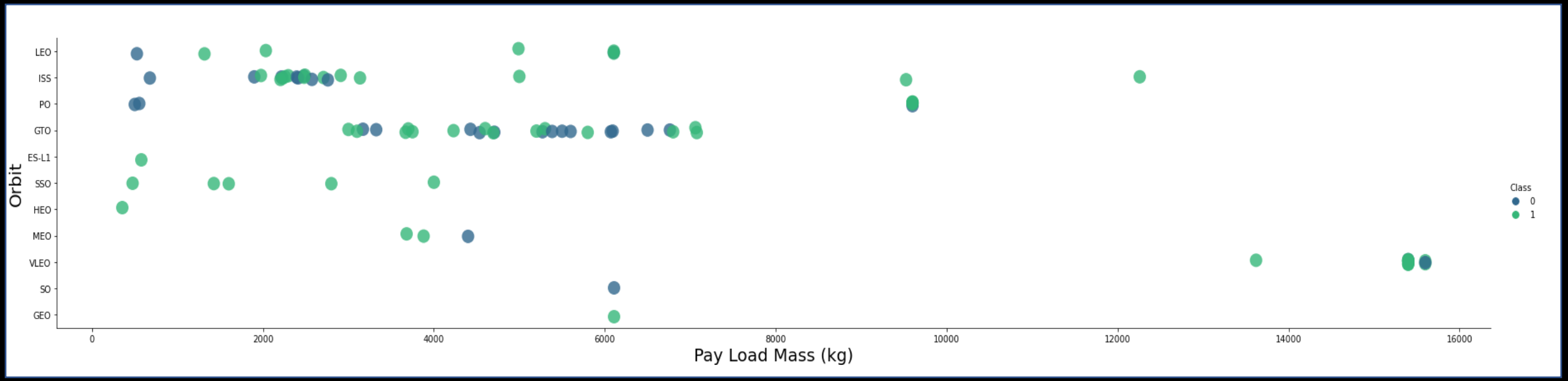
-Launch Orbit preferences changed over Flight Number.

-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



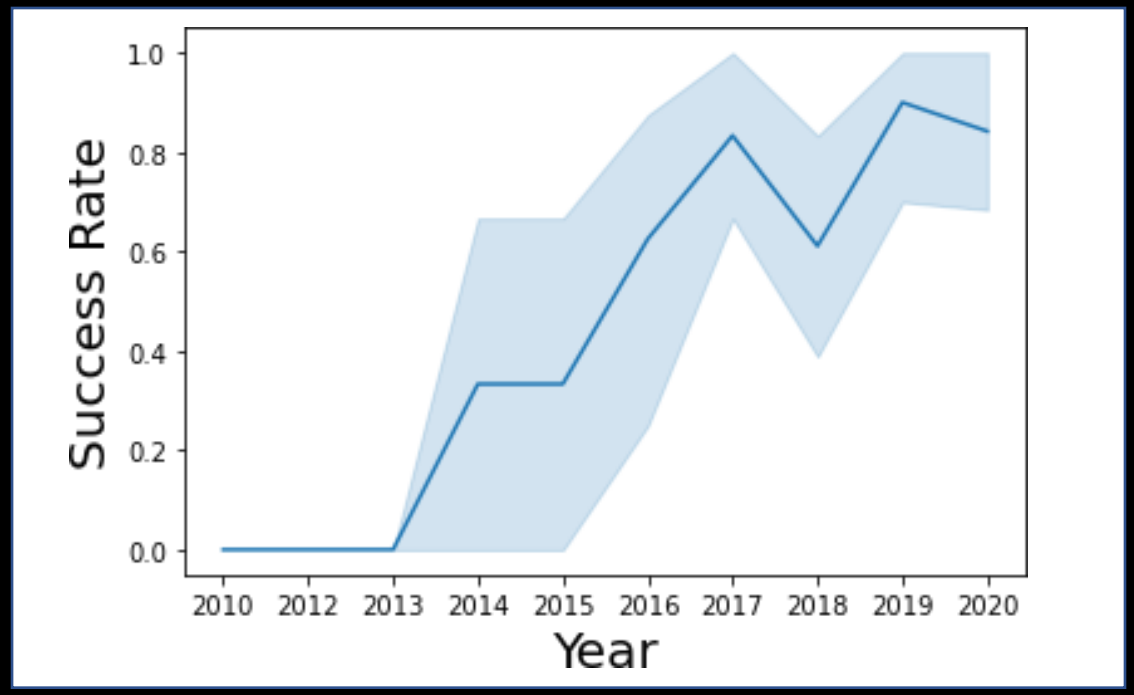
Green = successful launch, Blue = unsuccessful launch.

-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 higher payloads

Launch Success Yearly Trend



95% confidence interval
(light blue shading)

-Success generally increases over time since 2013 (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

-SQL 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 likely the previous name.

-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 SQL query sums the total payload mass (kg) where NASA was the customer.

-CRS: 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 lower end

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 successful ground pad landing wasn’t until the end of 2015.

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

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 99% of the time.

-This means that most of the landing failures are intended.

-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.

-All are of the F9 B5 B10xx.x variety.

-This likely indicates payload mass correlates with the booster version that used in launch.

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

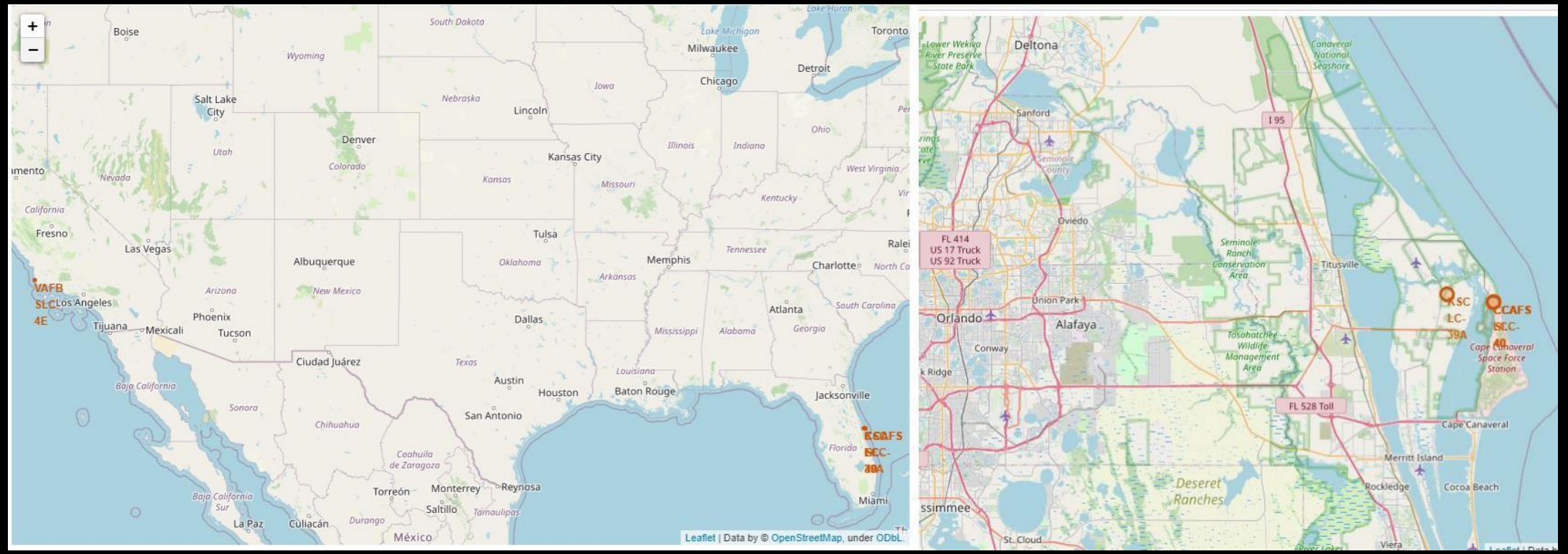
-This query returns a list of successful landings from 2010-06-04 to 2017-03-20

-Two types of successful landing outcomes: drone ship and ground pad landings.

-There were 8 successful landings during this time period

Interactive Map with Folium

Launch Site Locations

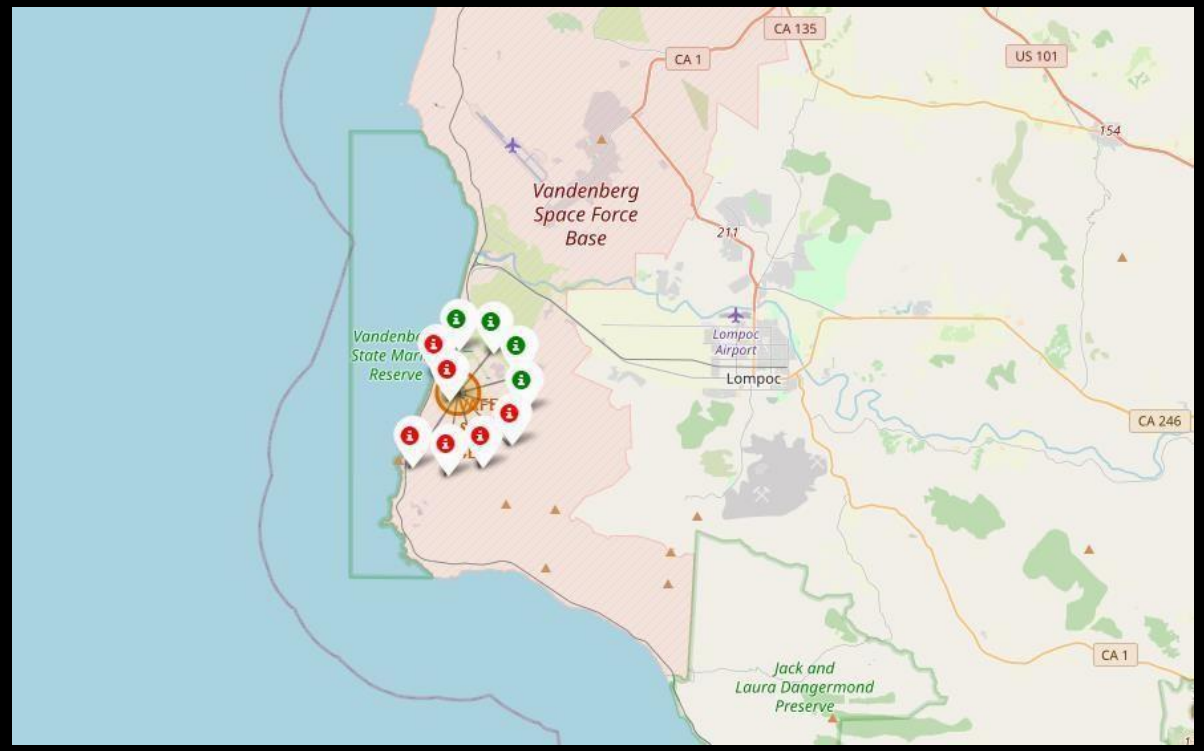


-The left map marks all launch sites in the US.

-The right map marks the two Florida launch sites

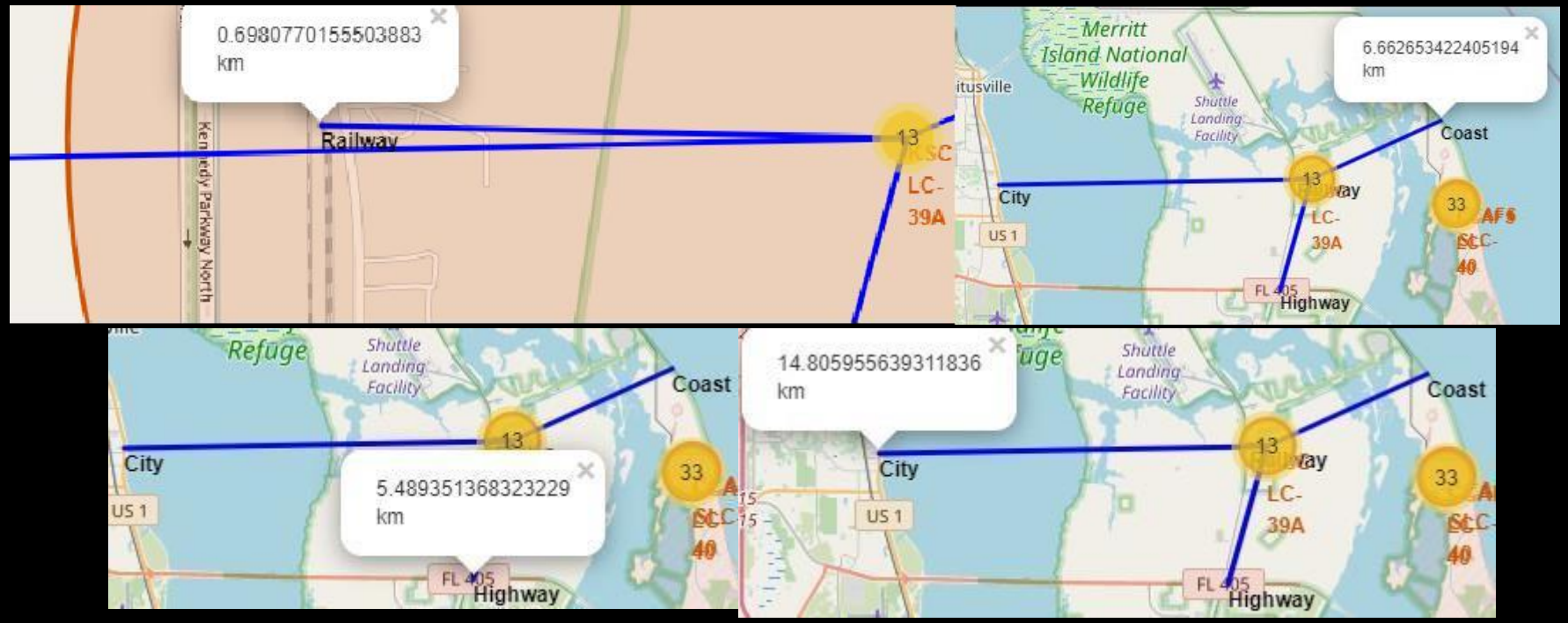
-All launch sites are near the ocean. (less risk of landing on buildings/populated areas!)

Color-Coded Launch Markers



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

Key Location Proximities



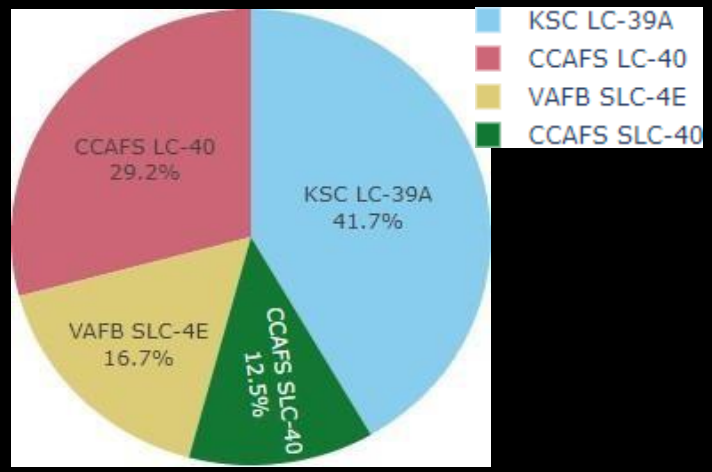
-Launch sites are very close to railways for large supply transportation.

-Launch sites are close to highways for human and other supply transportation.

-Launch sites are also close to coasts and relatively far from cities so that launch failures have a less risk of hitting populated areas/cities

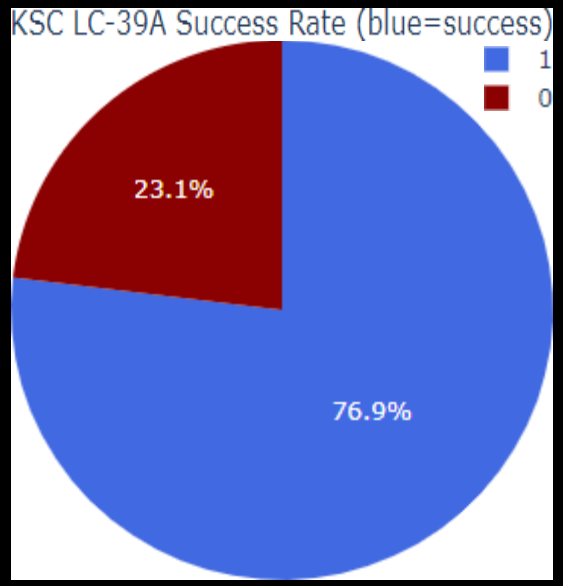
Build a Dashboard with Plotly Dash

Successful Launches Across Launch Sites



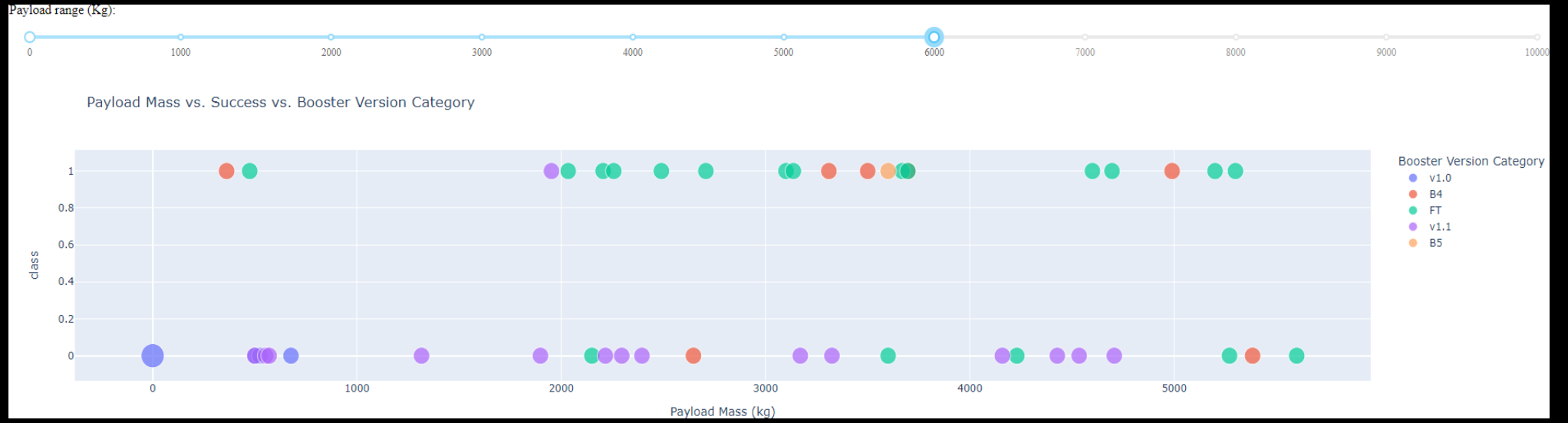
-The distribution of successful landings across all launch sites.
-CCAFS LC-40 is the old name of CCAFS SLC-40
-CCAFS and KSC have an equal amount of successful landings,
-Majority of the successful landings were performed before the name change.
-VAFB has the least amount of successful landings.
-VAFB had smaller sample
-Harder to launch in the west coast?

Highest Success Rate Launch Site



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

Payload Mass vs. Success vs. Booster Version Category



-Plotly dashboard has a Payload range selector.

-Class indicates 1 for successful landing and 0 for failure.

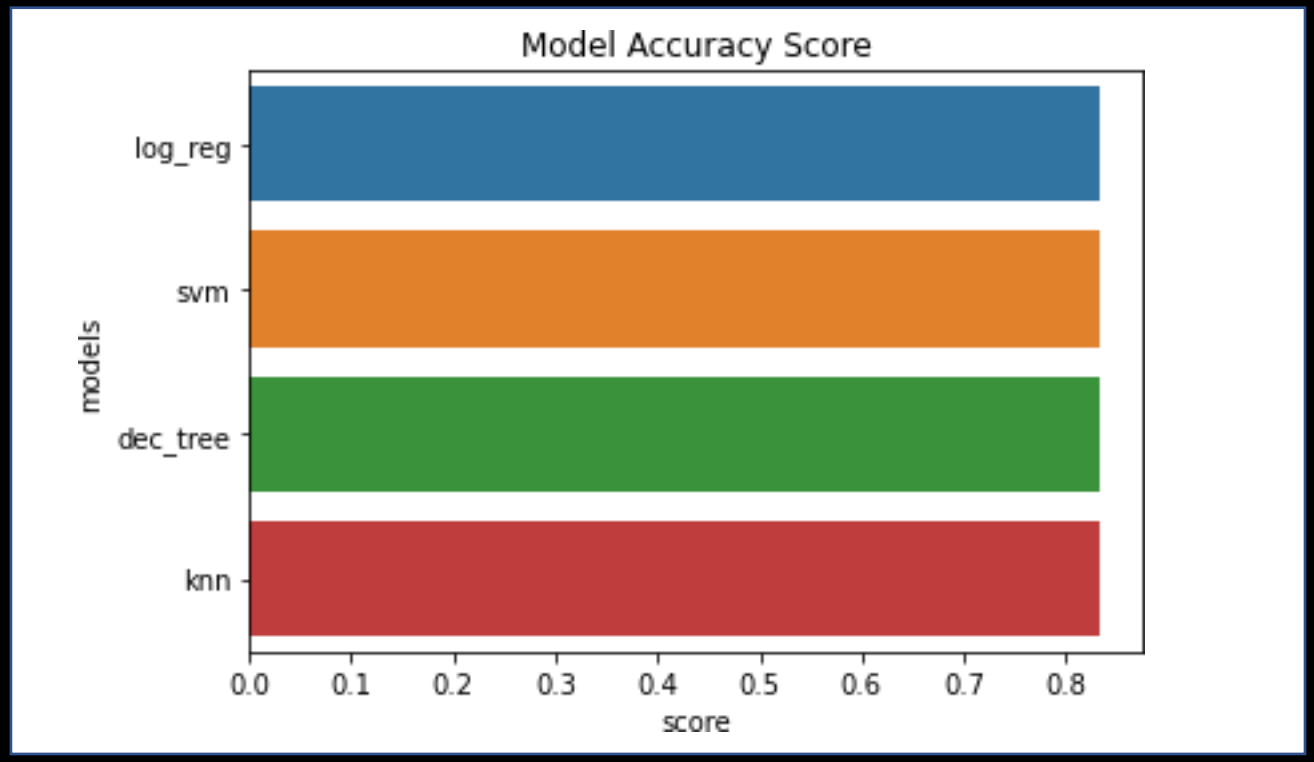
-Scatter plot accounts for booster version category in color and number of launches

-In range of 0-6000, 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



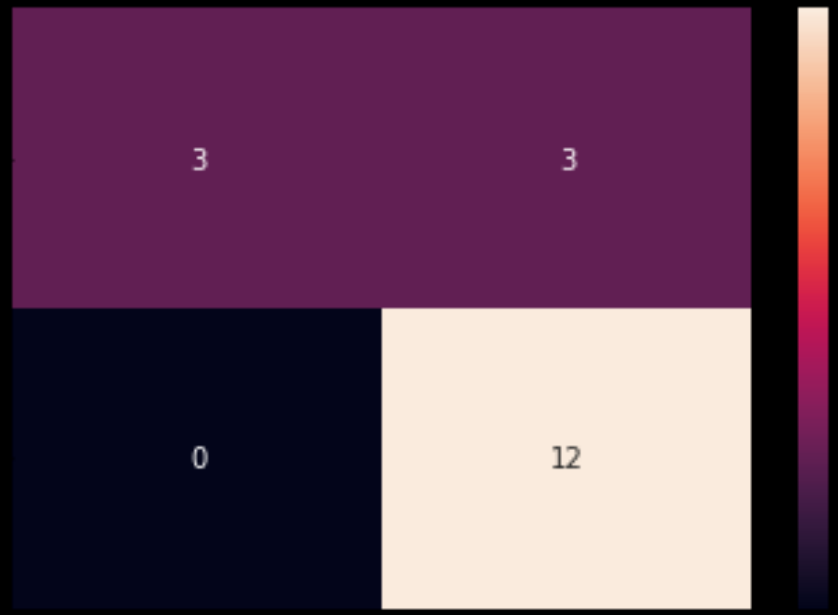
-All predictive models had the same accuracy on the test data (83.33% accuracy.)

-Test size is small, only a sample size of 18.

-This can cause large variance in accuracy results, such as those in Decision Tree Classifier model in repeated runs.

-We 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 overpredict successful landings.

CONCLUSION

-Our task: to develop a prediction model for Space Y to compete with Space X

-The goal of the 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 Plotly dashboard for visualization

-Created a machine learning model with an accuracy of 83%

-SpaceY can use this model to predict whether a launch will have a successful Stage 1 landing before actual launch to determine better business decisions of executing launch.

-HOWEVER, more data should be collected to further test and determine which prediction model has greatest accuracy.

Thank you!