Applied Data Science Capstone Project

Nan Yao

Agenda


Executive Summary
Introduction
Methodology, ED&A & Prediction
Interactive Map with Folium
Results & Summary

Executive Summary

- With using data collected from public SpaceX API and SpaceX Wikipedia page, we worked this project to predict Space X Falcon 9 first stage landing successful rate and this project covered data collection, exploration, visualization, machine learning modeling, folium maps, and dashboards.

- 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 is the only private company ever to return a spacecraft from low-earth orbit, which it first accomplished in December 2010. In this capstone project, we will predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars whereas other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage. Therefore if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used if an alternate company wants to bid against SpaceX for a rocket launch.

Data Collection and Wrangling

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.

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

```
launch_dict = {'FlightNumber': list(data['flight_number']),
'Date': list(data['date']),
'BoosterVersion':BoosterVersion,
'PayloadMass':PayloadMass,
'Orbit':Orbit,
'LaunchSite':LaunchSite,
'Outcome':Outcome,
'Flights':Flights,
'GridFins':GridFins,
'Reused':Reused,
'Legs':Legs,
'LandingPad':LandingPad,
'Block':Block,
'ReusedCount':ReusedCount,
'Serial':Serial,
'Longitude': Longitude,
'Latitude': Latitude}
```

FlightNumber	Date	BoosterVersion	PayloadMass	Orbit	LaunchSite	Outcome	Flights	GridFins	Reused	Legs	LandingPad	Block
1	2006-03-24	Falcon 1	20.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN
2	2007-03-21	Falcon 1	NaN	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN
4	2008-09-28	Falcon 1	165.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN
5	2009-07-13	Falcon 1	200.0	LEO	Kwajalein Atoll	None None	1	False	False	False	None	NaN

EDA and Interactive Visual Analytics

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is help to find some patterns in the data and determine what would be the label for training supervised models.

Calculate the number of launches on each site:

```
# Apply value_counts() on column LaunchSite
df["LaunchSite"].value_counts()
```

```
CCAFS SLC 40    55
KSC LC 39A      22
VAFB SLC 4E     13
Name: LaunchSite, dtype: int64
```

Calculate the number and occurrence of each orbit:

```
# Apply value_counts on Orbit column
df["Orbit"].value_counts()
```

```
GTO      27
ISS      21
VLEO     14
PO        9
LEO       7
SSO       5
MEO       3
SO        1
HEO       1
ES-L1     1
GEO       1
Name: Orbit, dtype: int64
```

Exploratory Data Analysis using SQL

With using SQL in Python and IBM DB2,

1. Queried using SQL Python integration.
2. Queries were made to get a better understanding of the dataset.
3. Queried information about launch site names, mission outcomes, various pay load sizes of customers and booster versions, and landing outcomes

Interactive Map with Folium

With using Folium in Python,

1. Folium maps mark Launch Sites, successful and unsuccessful landings, and a proximity example to key locations: Railway, Highway, Coast, and City.
2. This allows us to understand why launch sites may be located where they are. Also visualizes successful landings relative to location.

Space X Falcon 9 First Stage Landing Prediction

With using Machine Learning Classification Models,

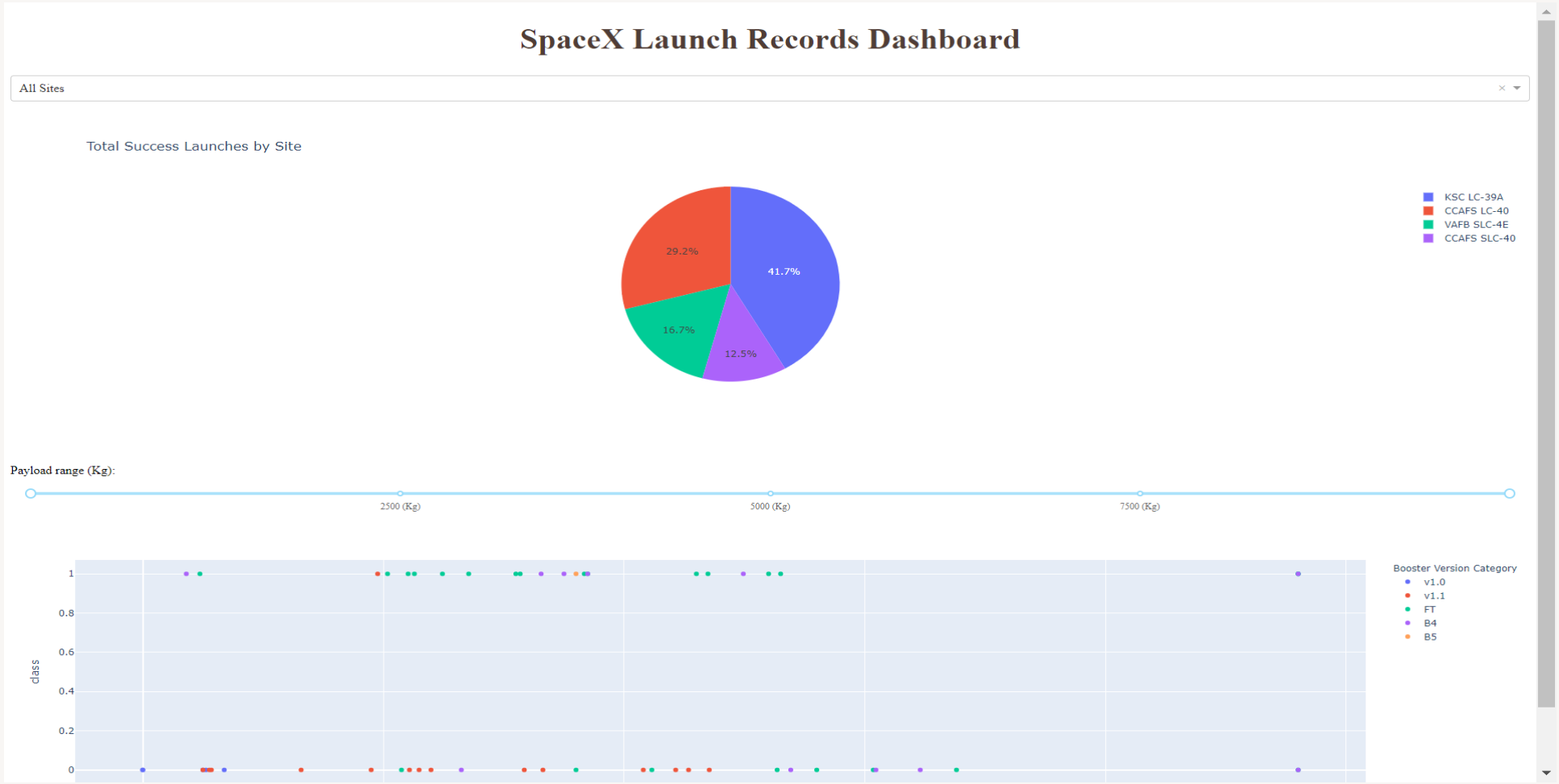
Perform exploratory Data Analysis and determine Training Labels

1. Create a column for the class
2. Standardize the data
3. Split into training data and test data

Find best Hyperparameter for SVM, Classification Trees and Logistic Regression

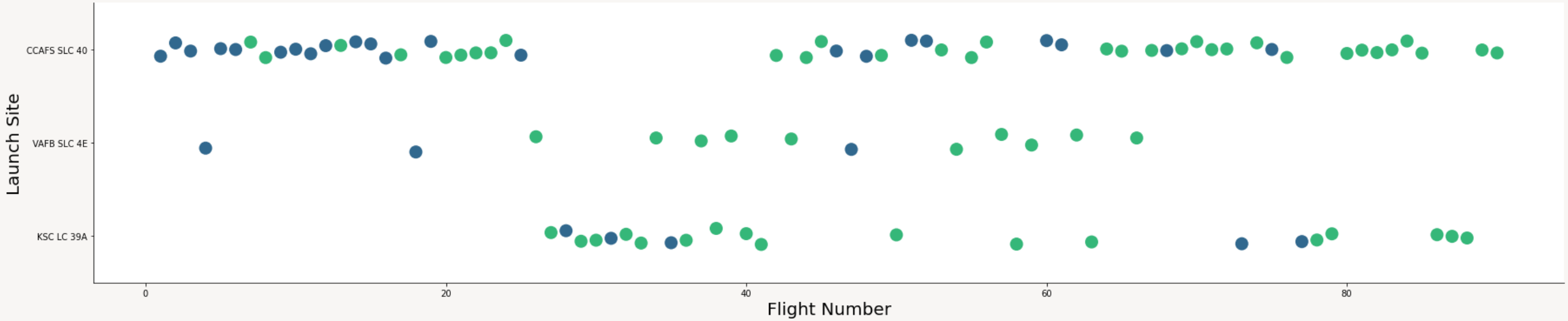
1. Find the method performs best using test data

Results


SpaceX Launch Records Dashboard
All Sites
Total Success Launches by Site
29.2%
41.7%
16.7%
12.5%
KSC LC-39A
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40
Payload range (Kg):
2500 (Kg)
5000 (Kg)
7500 (Kg)
class
1
0.8
0.6
0.4
0.2
0
Booster Version Category
v1.0
v1.1
FT
B4
B5

EDA with Visualization Results

Flight Number vs. Launch Site

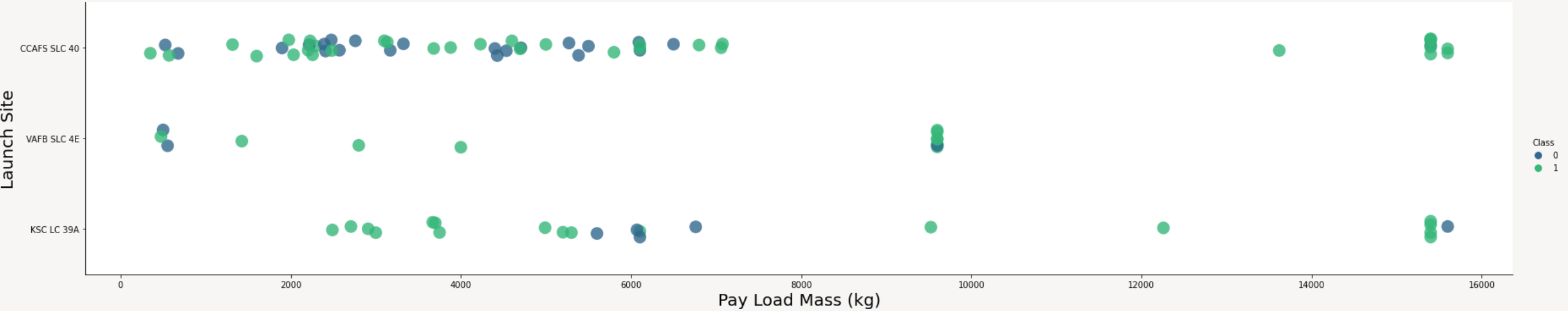


Green indicates successful launch; Blue indicates unsuccessful launch.

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.

EDA with Visualization Results

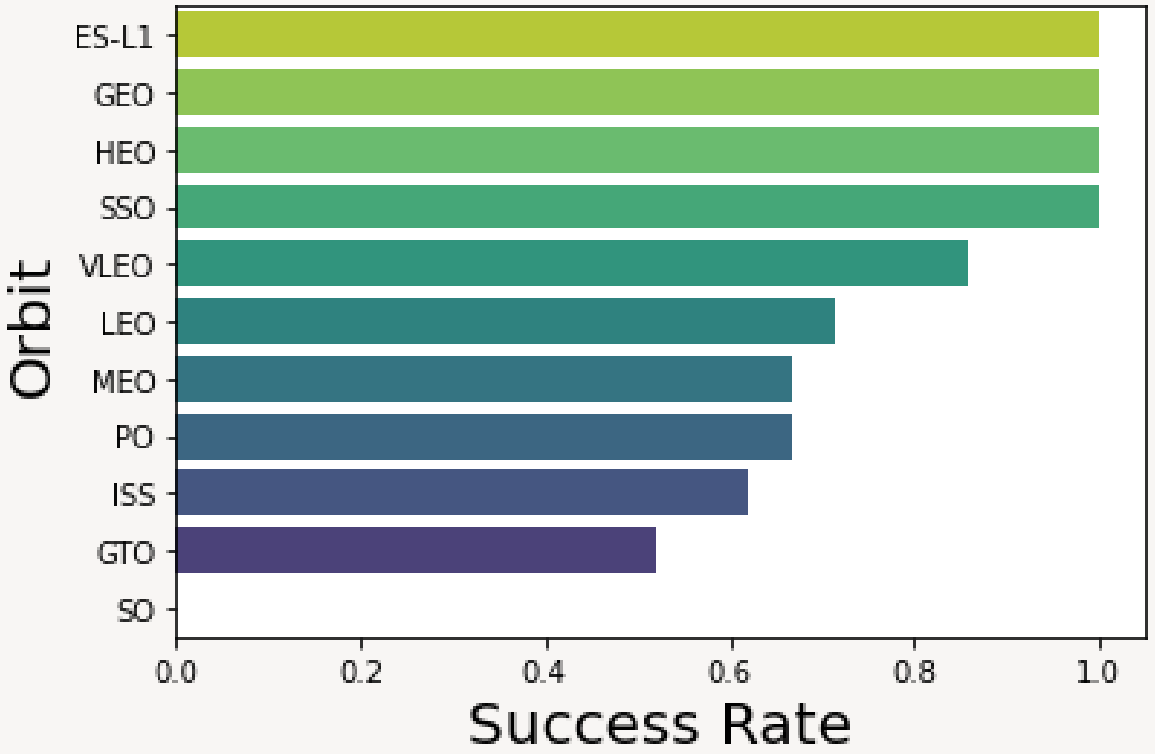
Payload vs. Launch Site



Green indicates successful launch; Blue indicates unsuccessful launch.

EDA with Visualization Results

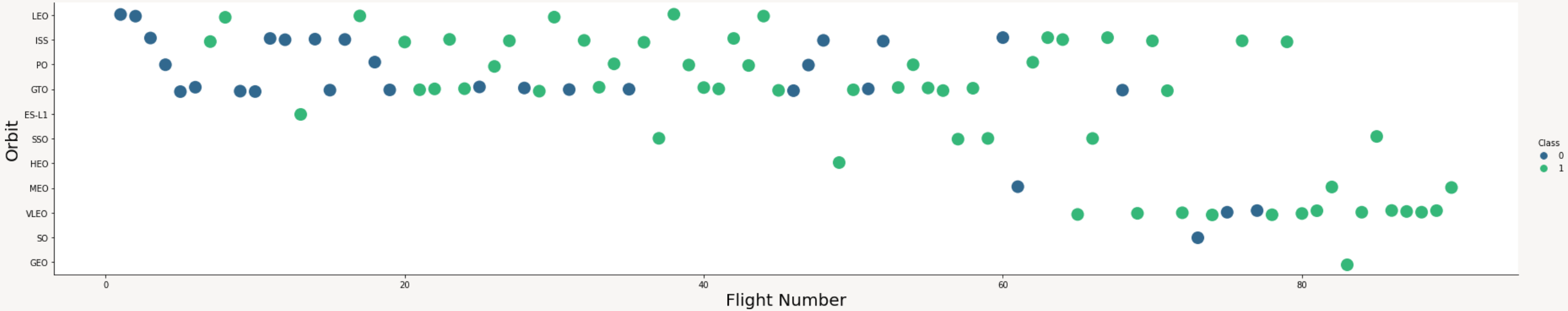
Success rate vs. Orbit type



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

EDA with Visualization Results

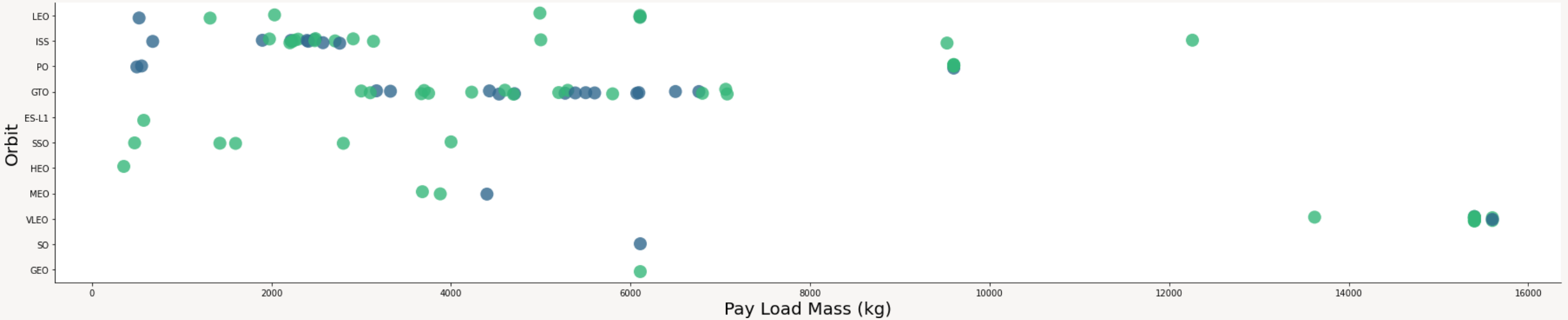
Flight Number vs. Orbit type



Green indicates successful launch; Blue indicates unsuccessful launch.

EDA with Visualization Results

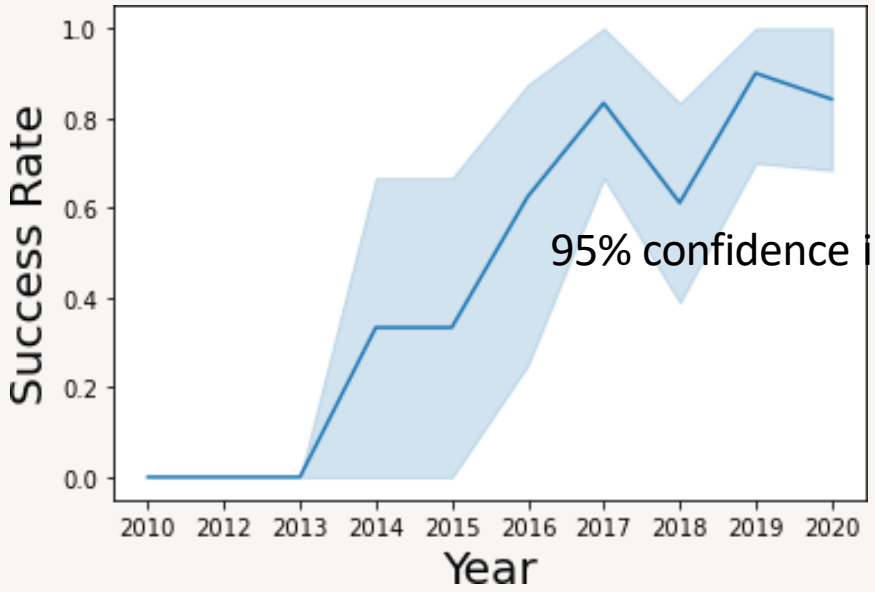
Payload vs. Orbit type



Green indicates successful launch; Blue indicates unsuccessful launch.

EDA with Visualization Results

Launch Success Yearly Trend



95% confidence interval (light blue shading)

95% confidence interval (light blue shading)

EDA with SQL Results

All Unique Launch Site Names

%sql select DISTINCT LAUNCH_SITE from SPACEXDATASET

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

EDA with SQL Results

List 5 records where launch sites begin with the string 'CCA'

%sql select * from SPACEXDATASET where launch_site like 'CCA%' limit 5

DATE	time_utc_	booster_version	launch_site	payload	payload_mass_kg_	orbit	customer	mission_ou
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	S
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	S
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	S
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	S
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	S

EDA with SQL Results

Display the total payload mass carried by boosters launched by NASA (CRS)

```
%sql select sum(payload_mass__kg_) as sum from SPACEXDATASET where customer like 'NASA (CRS)'
```

SUM
45596

EDA with SQL Results

Display average payload mass carried by booster version F9 v1.1

%sql select avg(payload_mass__kg_) as Average from SPACEXDATASET where booster_version like 'F9 v1.1%'

average
2534

EDA with SQL Results

List the date when the first succesful landing outcome in ground pad was acheived.

%sql select min(date) as Date from SPACEXDATASET where mission_outcome like 'Success'

DATE
2010-06-04

EDA with SQL Results

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

%sql select booster_version from SPACEXDATASET where (mission_outcome like 'Success')
AND (payload_mass__kg_ BETWEEN 4000 AND 6000) AND (landing__outcome like 'Success (drone ship)')

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

EDA with SQL Results

List the total number of successful and failure mission outcomes

%sql SELECT mission_outcome, count(*) as Count FROM SPACEXDATASET GROUP by mission_outcome ORDER BY mission_outcome

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

EDA with SQL Results

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

```
maxm = %sql select max(payload_mass__kg_)
from SPACEXDATASET
maxv = maxm[0][0]
%sql select booster_version from
SPACEXDATASET where
payload_mass__kg_=(select
max(payload_mass__kg_) from SPACEXDATASET)
```

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

EDA with SQL Results

List the records which will display the month names, failure landing_outcomes in drone ship ,booster versions, launch_site for the months in year 2015

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

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

EDA with SQL Results

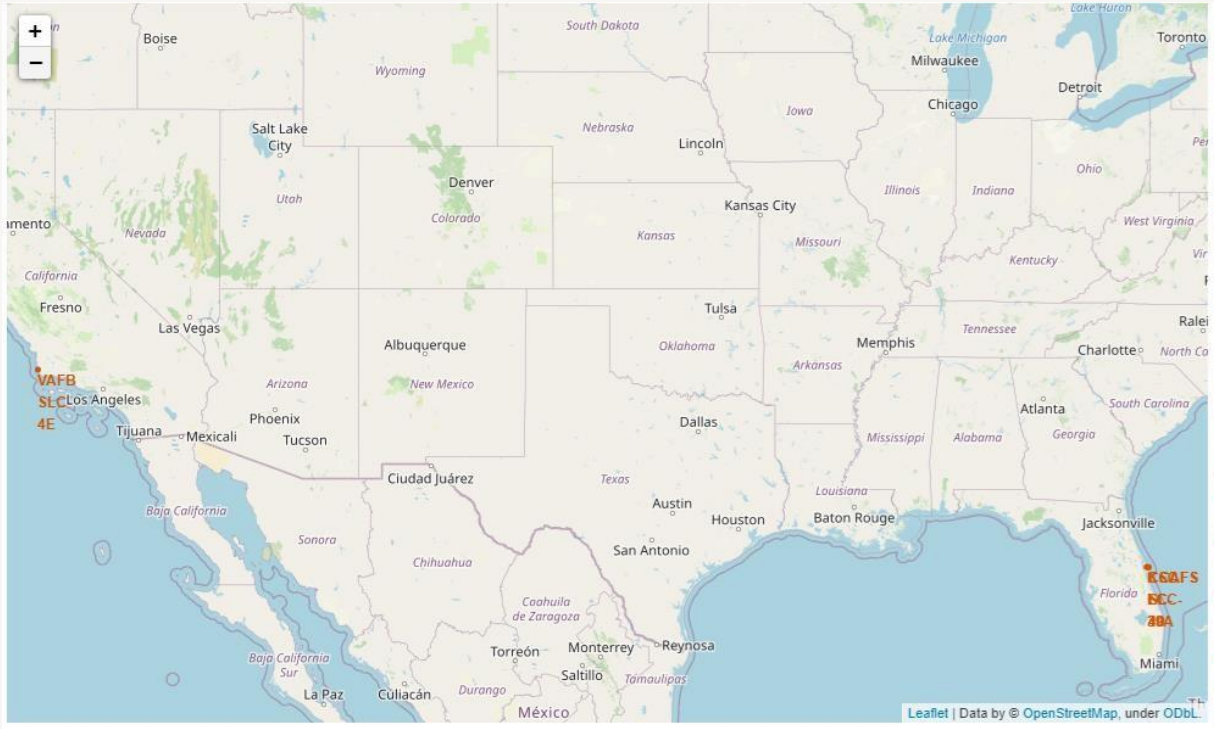
Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.

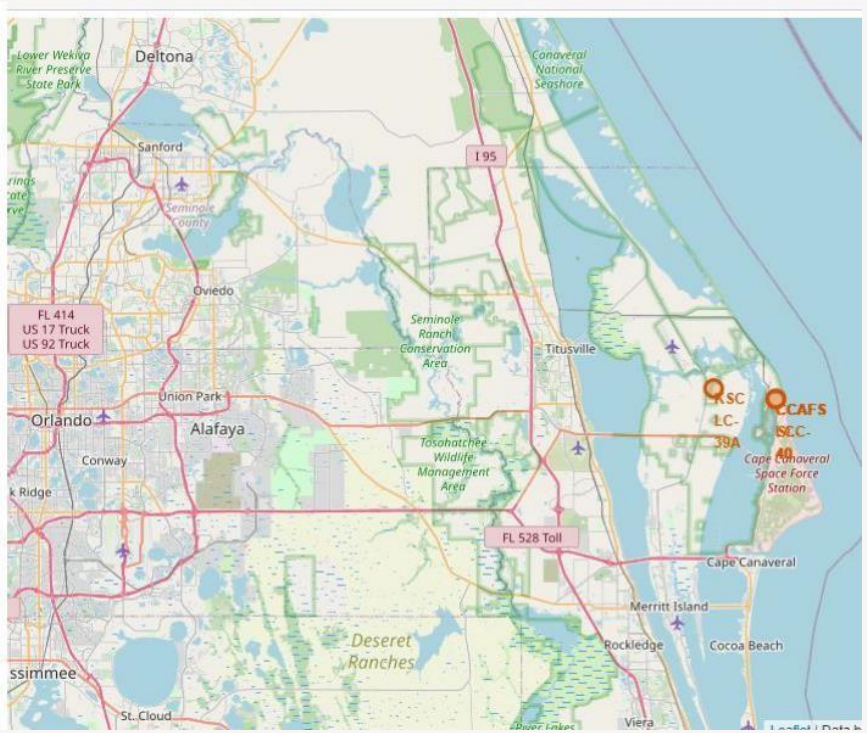
```
%sql select landing__outcome, count(*) as count from SPACEXDATASET
where Date >= '2010-06-04' AND Date <= '2017-03-20'
GROUP by landing__outcome ORDER BY count Desc.
```

landing__outcome	COUNT
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Interactive Map with Folium Results

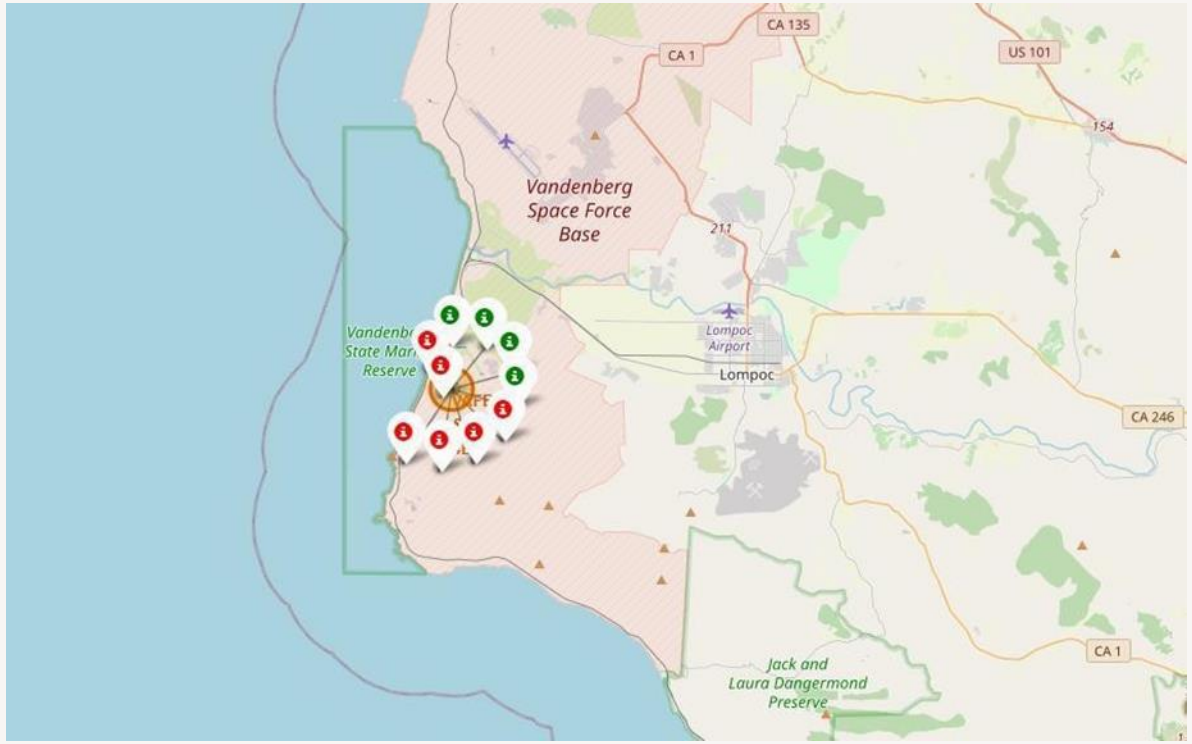
Launch Site





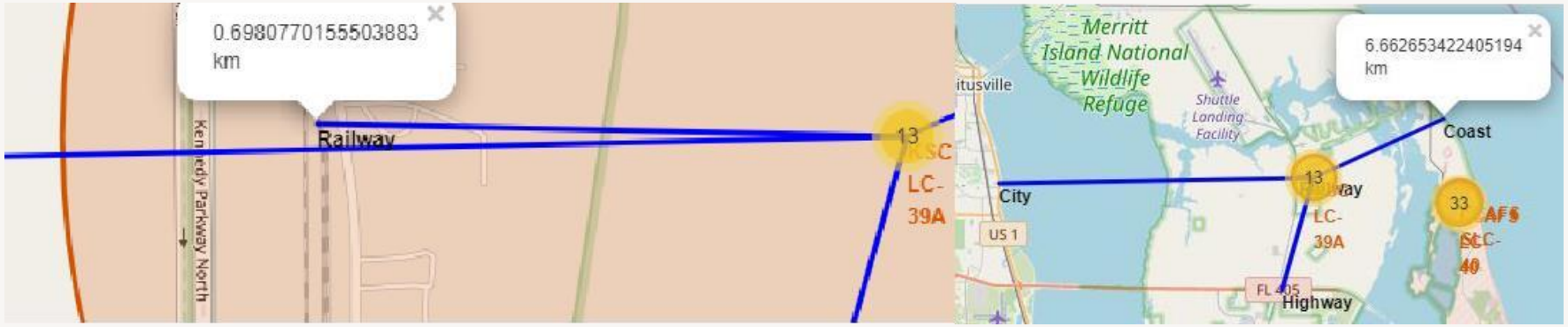
Interactive Map with Folium Results

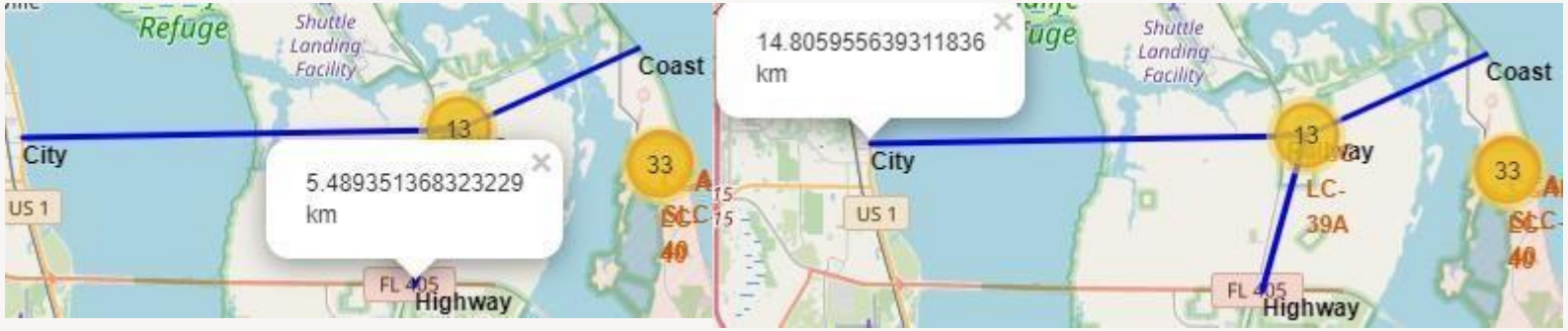
Color-Coded Launch Markers



Interactive Map with Folium Results

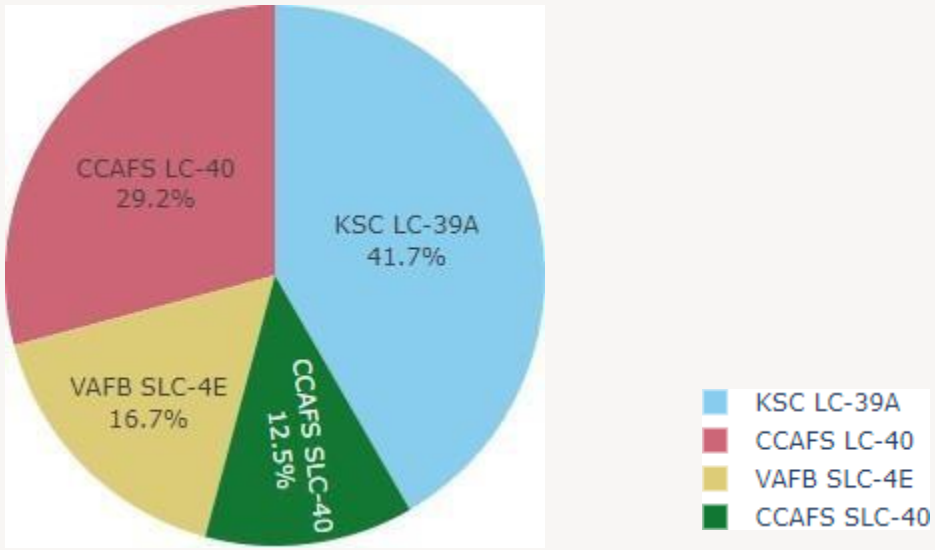
Key Location Proximities





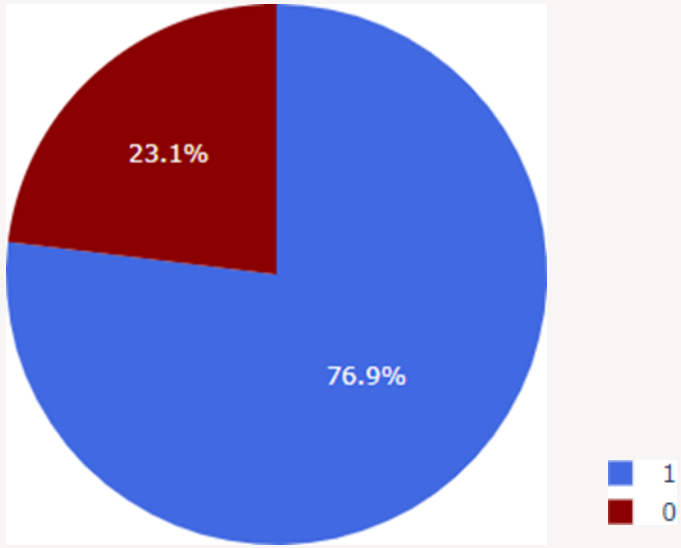
Plotly Dash dashboard Results

Successful Launches Across Launch Site



Plotly Dash dashboard Results

Highest Success Rate Launch Site



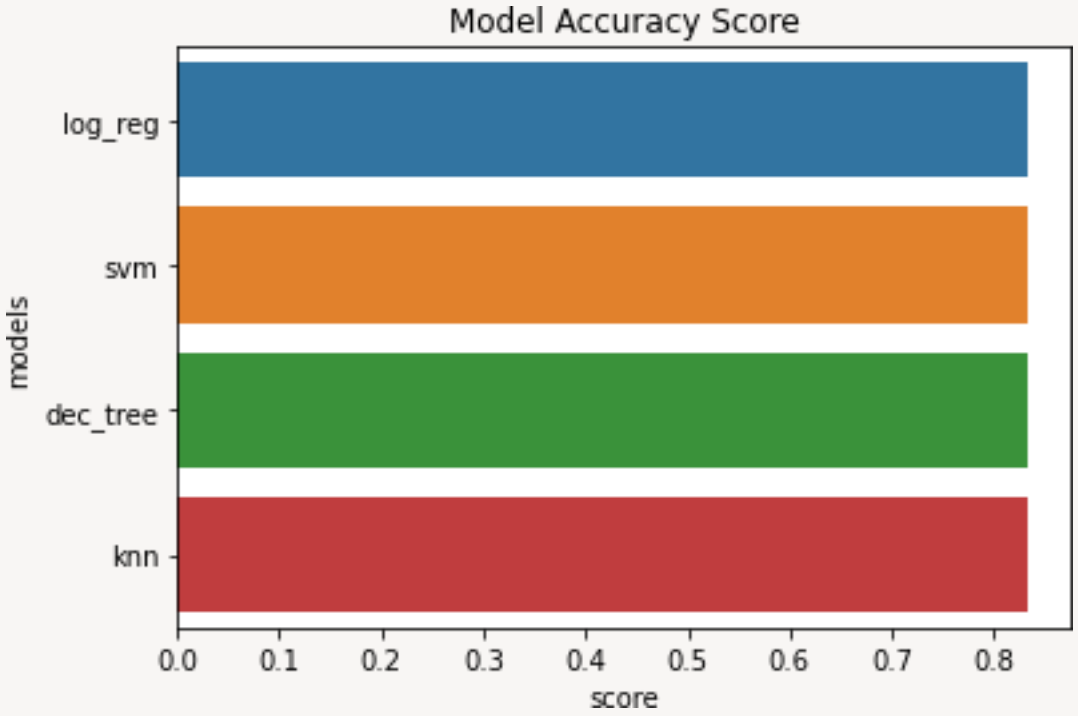
KSC LC-39A Success Rate (blue=success)

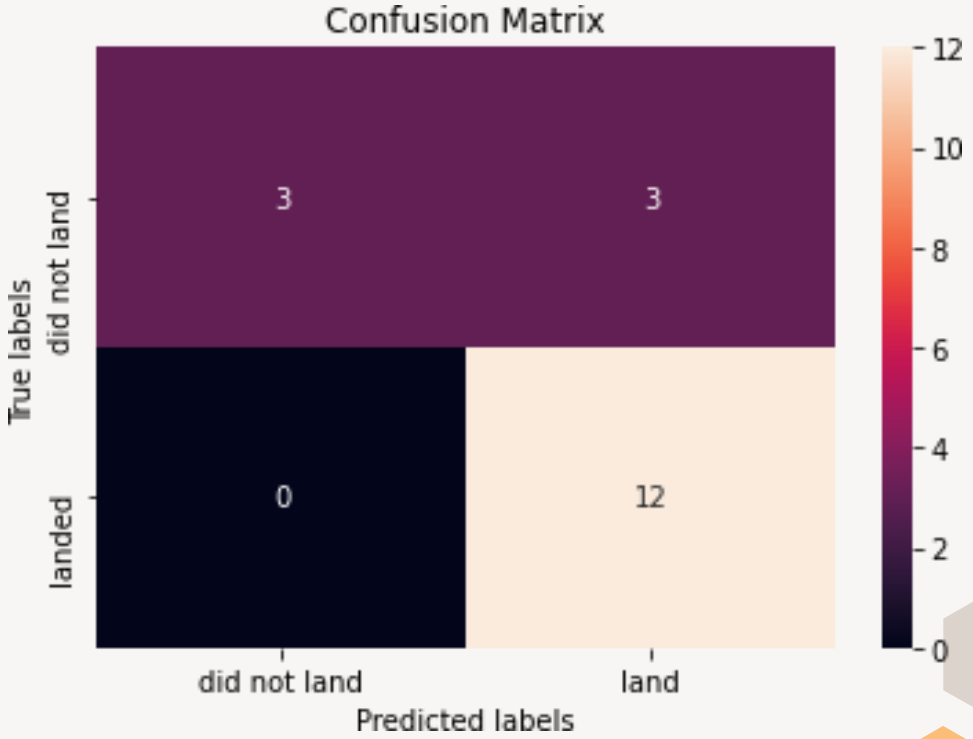
Plotly Dash dashboard Results

Payload Mass vs. Success vs. Booster Version Category

Predictive Analysis (Classification) Results

Classification Accuracy & Confusion Matrix





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

Conclusion

Our task is 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

If possible more data should be collected to better determine the best machine learning model and improve accuracy

Thank you