IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mahmoud Elmallah
14/01/2025

Outline

Executive Summary

Collected data from public SpaceX API and SpaceX Wikipedia page. Created labels column ‘class’ which classifies successful landings. Explored data using SQL, visualization, folium maps, and dashboards. Gathered relevant columns to be used as features. Changed all categorical variables to binary using one hot encoding. Standardized data and used GridSearchCV to find best parameters for machine learning models. Visualize accuracy score of all 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

Background

- Commercial Space age is here
- Space X has best pricing ($62 million vs $165 million USD)
- Reuse the first stage of falcon 9
- Space Y compete with Space X

Problem

The first stage is one of the most expensive part of rocket. And it’s important to the company to reuse It. We will train a machine learning Model to predict successful stage 1 recovery

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Combined data from SpaceX API and SpaceX Wikipedia page
- Perform data wrangling
  - Classifying true landings as successful and unsuccessful otherwise
  - Determine the number for each orbit and launch site
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Use ML models by GridSearchCV

Data Collection

Data collection process involved a combination pf API requests from SpaceX and Web Scraping

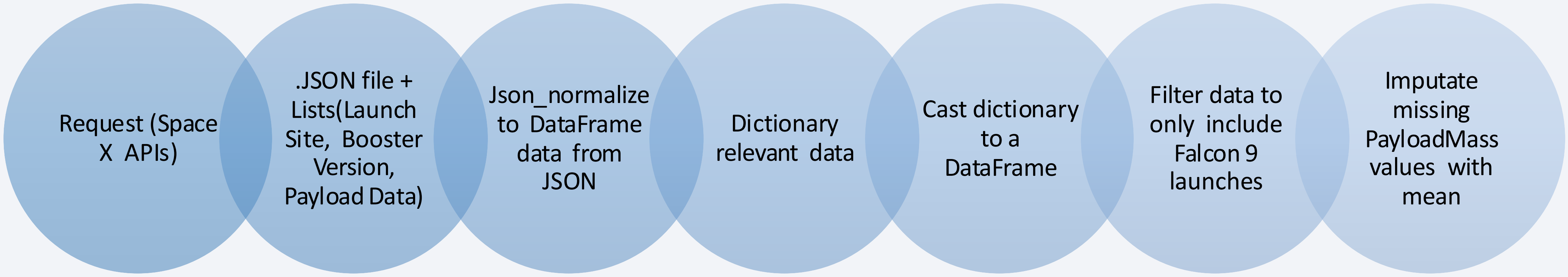
- Space X API Data Columns:

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

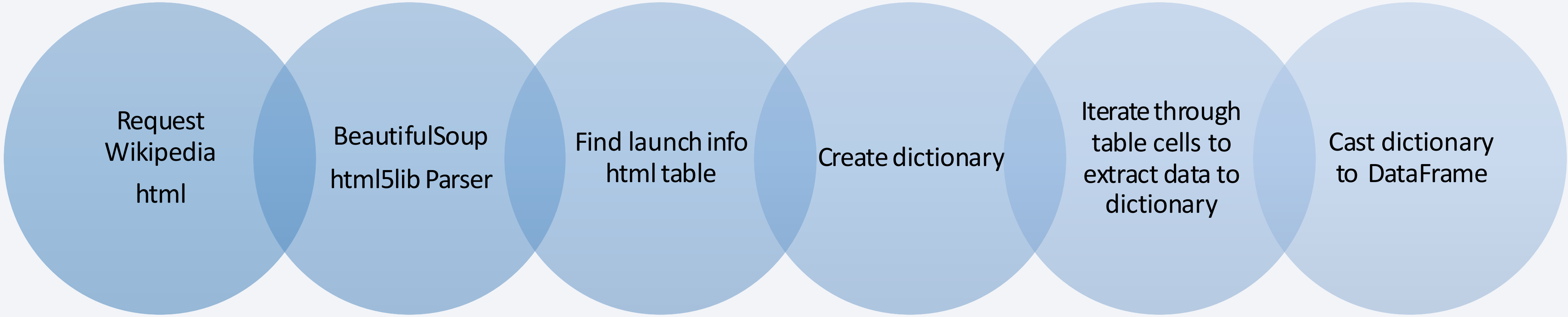
Data Collection – SpaceX API



GitHub URL:

https://github.com/Mahmoud-Mohamed-Almallah/Applied_Data_Science_IBM/blob/main/jupyter-labs-spacex-data-collection-api-v2.ipynb

Data Collection - Scraping



GitHub URL:

https://github.com/Mahmoud-Mohamed-Almallah/Applied_Data_Science_IBM/blob/main/jupyter-labs-webscraping.ipynb

Data Wrangling

Create a training label with landing outcomes where successful = 1 & failure = 0

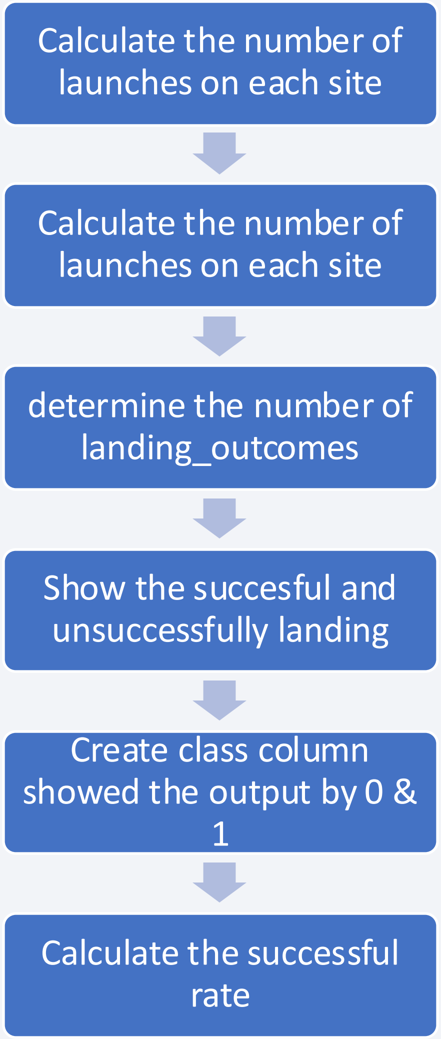
Outcome column has two components: ‘Mission Outcome’ ‘Landing Location’

New training label column ‘class’ with a value of 1 if ‘Mission Outcome’

is True and 0 otherwise.

True ASDS, True RTLS, & True Ocean – set to -> 1

None None, False ASDS, None ASDS, False Ocean, False RTLS – set to -> 0



GitHub URL:

https://github.com/Mahmoud-Mohamed-Almallah/Applied_Data_Science_IBM/blob/main/labs-jupyter-spacex-Data wrangling-v2.ipynb

EDA with Data Visualization

Exploratory Data Analysis 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

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

GitHub URL:

https://github.com/Mahmoud-Mohamed-Almallah/Applied_Data_Science_IBM/blob/main/jupyter-labs-eda-dataviz-v2.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,

GitHub URL:

https://github.com/Mahmoud-Mohamed-Almallah/Applied_Data_Science_IBM/blob/main/jupyter-labs-eda-sql-coursera_sqllite.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.

Made objects like Circles, markers and lines

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

GitHub URL:

https://github.com/Mahmoud-Mohamed-Almallah/Applied_Data_Science_IBM/blob/main/lab-jupyter-launch-site-location-v2.ipynb

Build a Dashboard with Plotly Dash

Dashboard includes a pie chart and a scatter plot.

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

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

The pie chart is used to visualize launch site success rate.

The scatter plot can help us see how success varies across launch sites, payload mass, and booster version category.

GitHub URL:

https://github.com/Mahmoud-Mohamed-Almallah/Applied_Data_Science_IBM/blob/main/spacex_dash_app.py

Predictive Analysis (Classification)

Split label column ‘Class’ from database

Fit and Transform Features using SandardScaler

Train_test_split data function

GridSearshCV (cv=10) to find optimal parametres

Use GridSearchCV on LogReg, SVM, Decision Tree and KNN models

Score model on split test set (train, test)

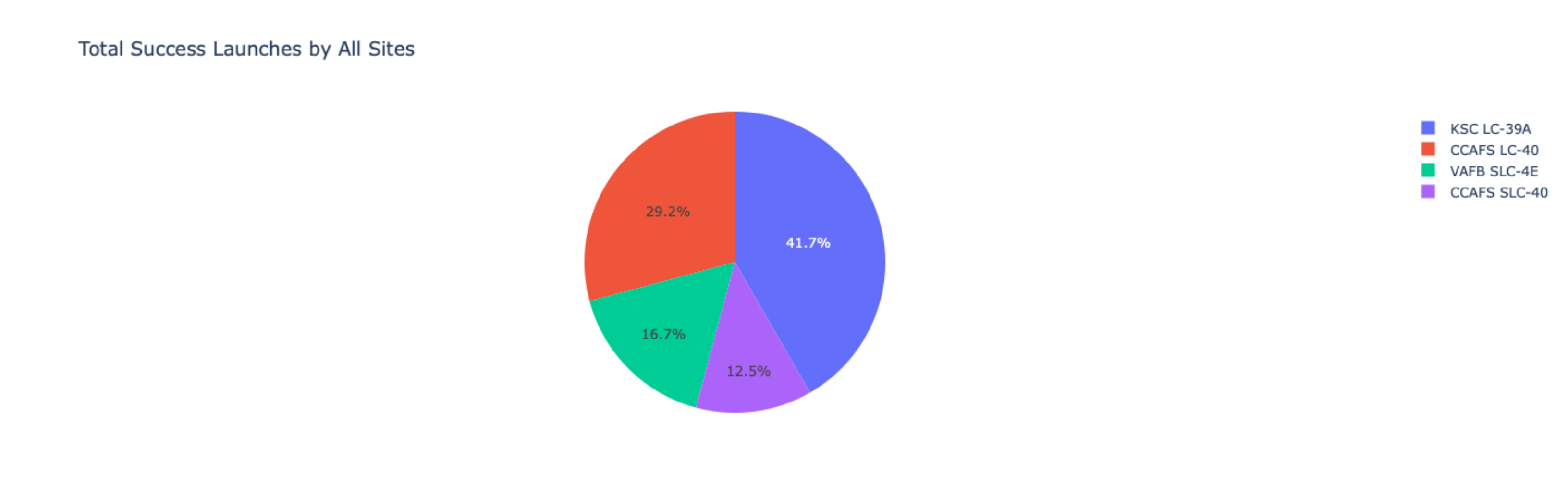
Confusion Matrix for all models

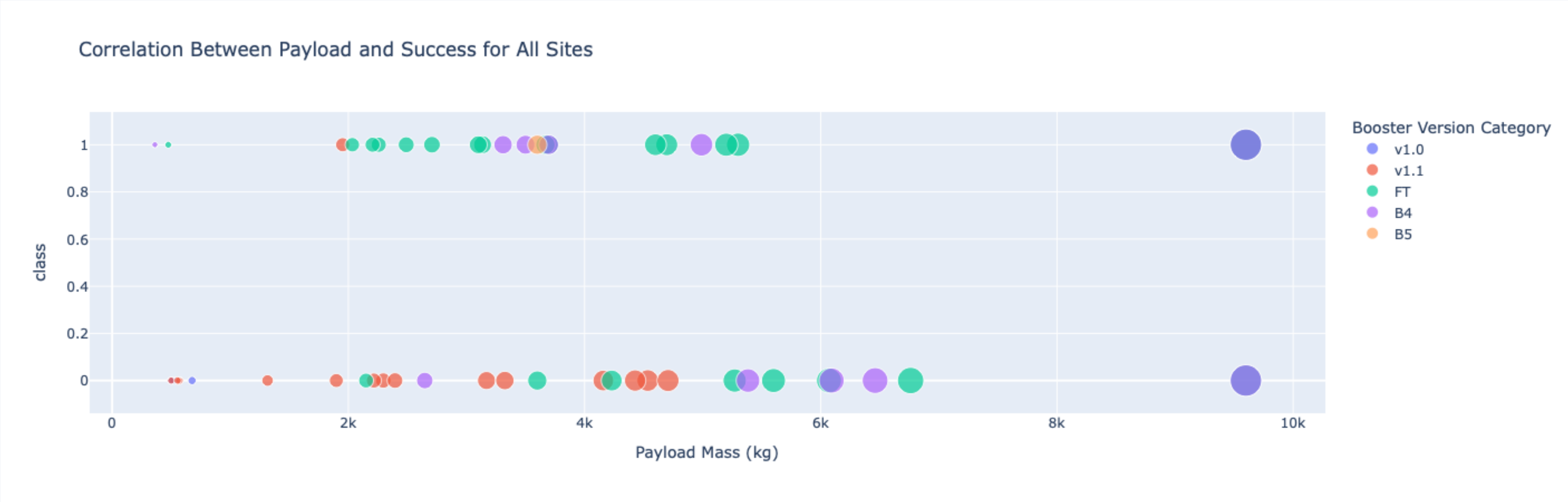
BarPlot to compare scores of models

GitHub URL:

https://github.com/Mahmoud-Mohamed-Almallah/Applied_Data_Science_IBM/blob/main/SpaceX-Machine-Learning-Prediction-Part-5-v1.ipynb

Results

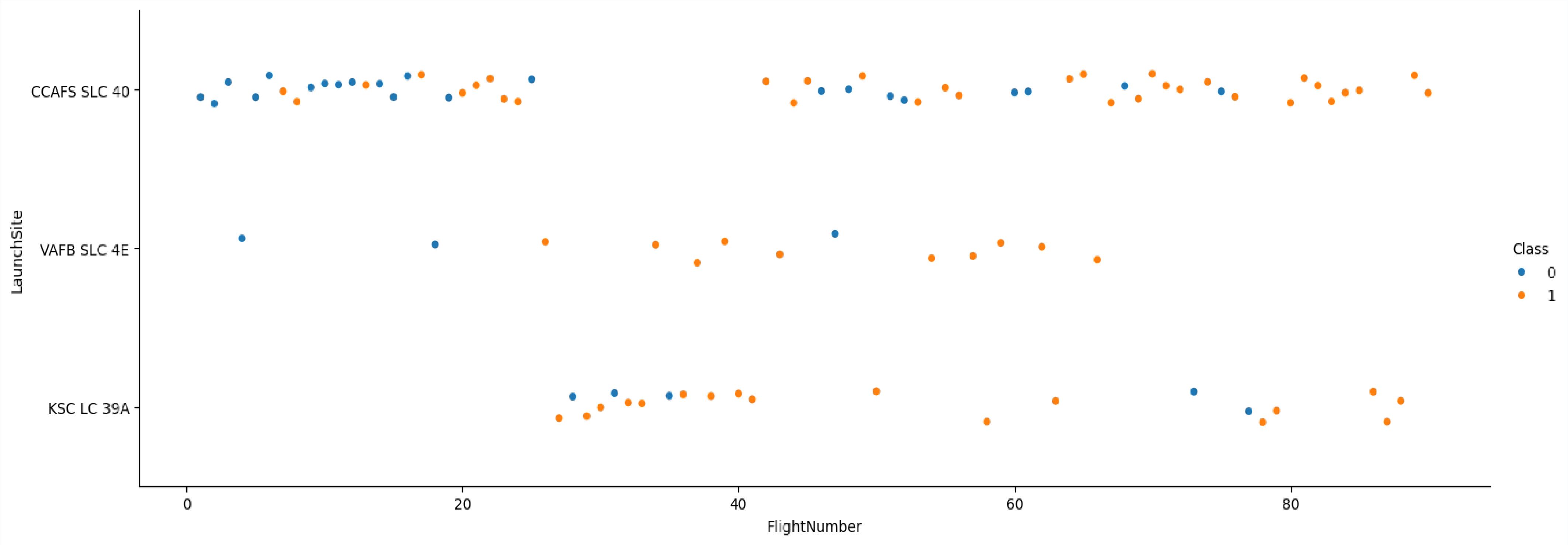

Total Success Launches by All Sites
41.7%
29.2%
16.7%
12.5%
KSC LC-39A
CCAFS LC-40
VAFB SLC-4E
CCAFS SLC-40


Correlation Between Payload and Success for All Sites
class
1
0.8
0.6
0.4
0.2
0
0
2k
4k
6k
8k
10k
Payload Mass (kg)
Booster Version Category
v1.0
v1.1
FT
B4
B5

Section 2

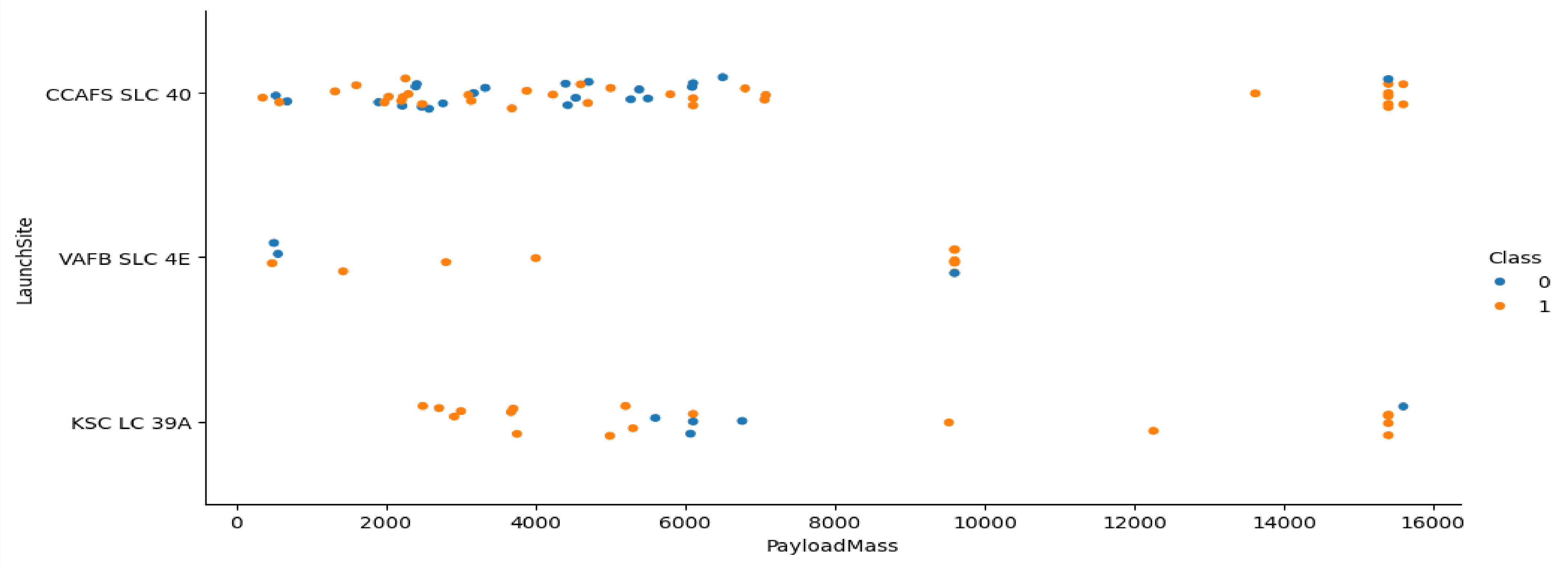
Insights drawn from EDA

Flight Number vs. Launch Site



CCAFS SLC 40 has more flights and most success launch
KSL LC 39A has the least number of flights but most of them are successful

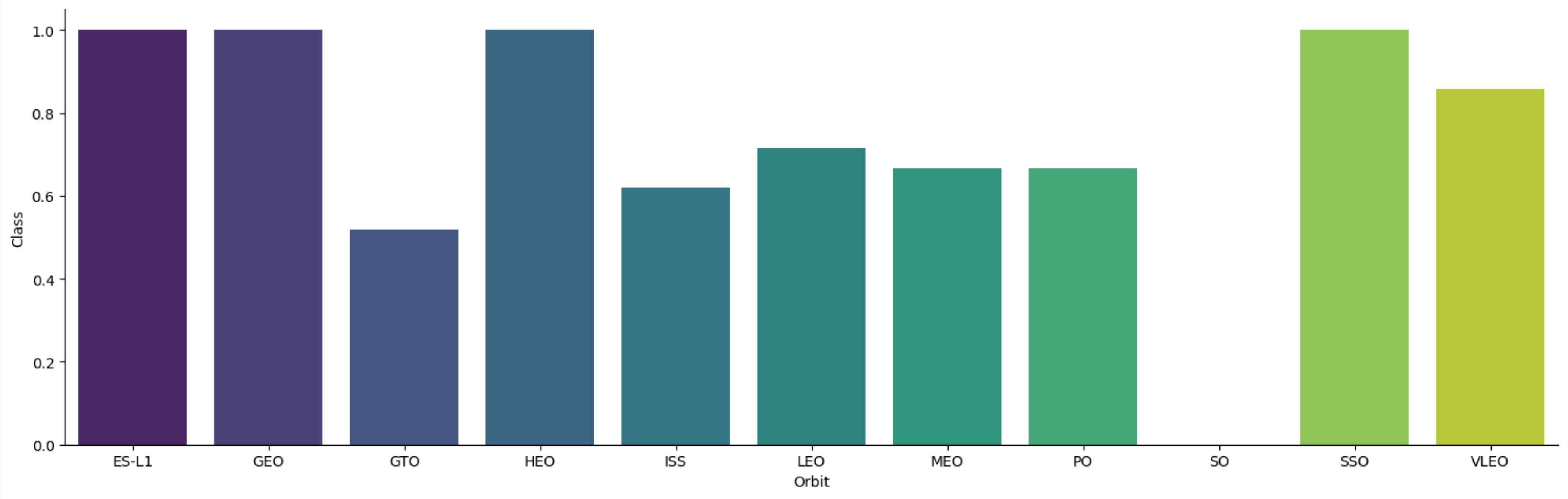
Payload vs. Launch Site



Successful launch = 1 Unsuccessful launch = 0

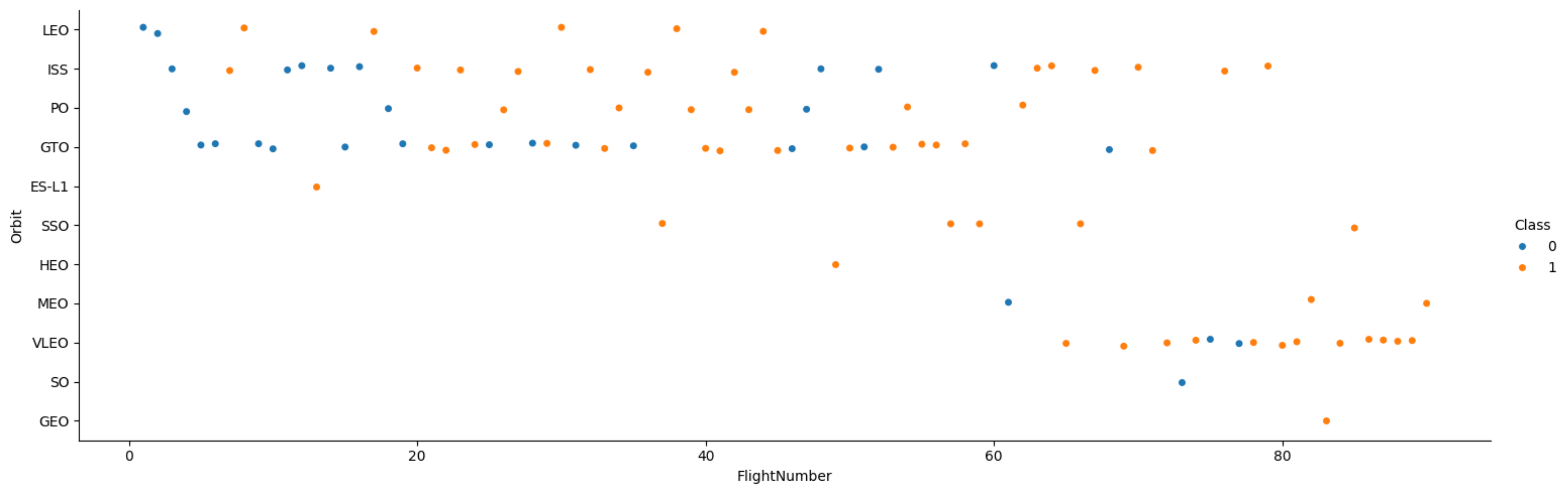
Payload mass appears to fall mostly between 0-6000 kg.
Different launch sites also seem to use different payload mass.

Success Rate vs. Orbit Type



- ES-L1 (1) , GEO (1) , HEO (1) have 100% success rate
- SSO (5) has 100% success rate
- VLEO (14) has more than 80% success rate
- GTO (27) has the around 50% success rate but largest sample
- SO (1) has 0% success rate

Flight Number vs. Orbit Type



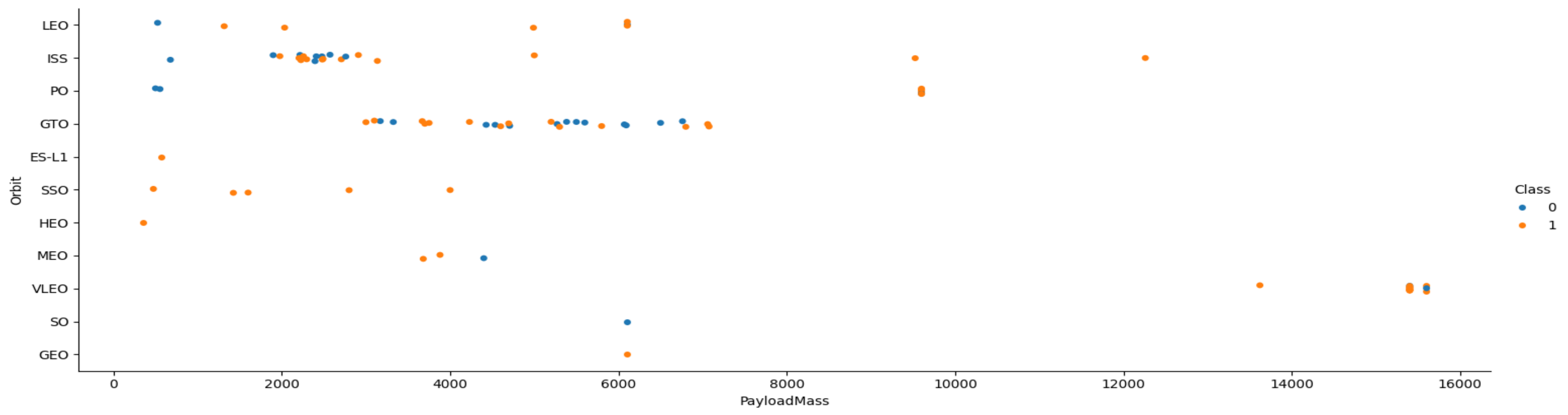
Successful launch = 1 Unsuccessful launch = 0

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\

Payload vs. Orbit Type



Successful launch = 1 Unsuccessful launch = 0

Payload mass seems to correlate with orbit

LEO and SSO seem to have relatively low payload mass

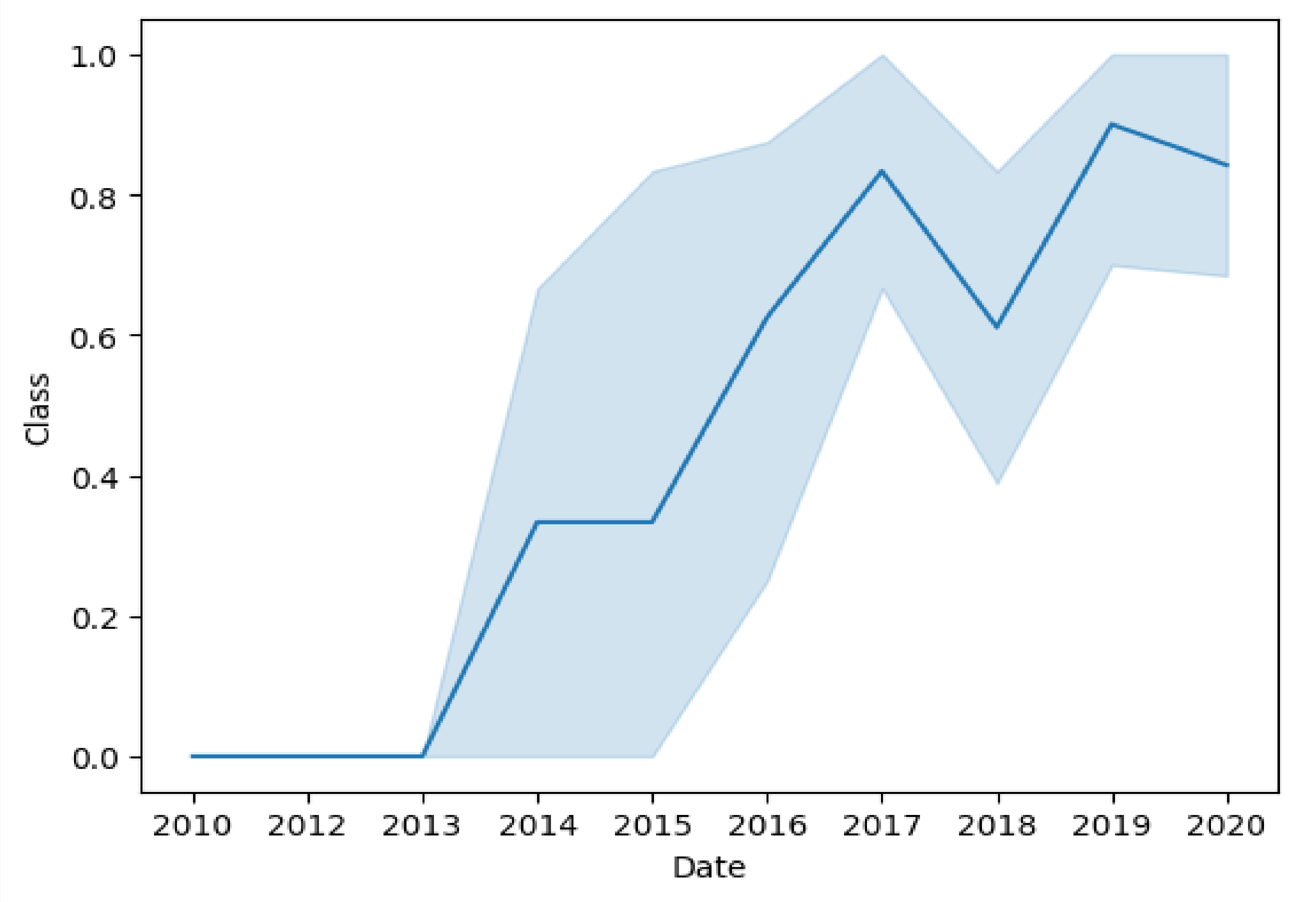
The other most successful orbit VLEO only has payload mass values in the higher end of the range

Launch Success Yearly Trend

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

Success in recent years at around 80%

the success rate since 2013 kept increasing till 2017 (stable in 2014) and after 2015 it started increasing



All Launch Site Names

Query unique launch site names from database.

CCAFS SLC-40 and CCAFSSLC-40 likely all represent the same launch site with data entry errors so I used distinct to show it only one.

CCAFS LC-40 was the previous name.

Likely only 3 unique launch_site values: CCAFS SLC-40, KSC LC-39A,VAFB SLC-4E

```
%sql SELECT distinct launch_site FROM SPACEXTABLE;
```

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

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

Launch Site Names Begin with 'CCA'

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

```
%sql select * from SPACEXTABLE where launch_site like 'CCA%' limit 5;
```

✓ 0.0s Python

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

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).

```
%%sql select sum(PAYLOAD_MASS__KG_) as total_payload_mass
from SPACEXTABLE where customer like 'NASA (CRS)';
✓ 0.0s

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

total_payload_mass
45596

Average Payload Mass by F9 v1.1

This query calculates the average payload mass or launches which used booster version F9 v1.1

Average payload mass of F9 1.1 is on the low end of our payload mass range

```
%%sql select avg(PAYLOAD_MASS__KG_) avg_payload_mass
from SPACEXTABLE where booster_version like '%F9 v1.1%';
```

✓ 0.0s

* sqlite:///my_data1.db
Done.

avg_payload_mass
2534.6666666666665

First Successful Ground Landing Date

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.

```
%%sql select min(date) as first_successful_landing
from SPACEXTABLE where Landing_Outcome = 'Success (ground pad)';
```

✓ 0.0s

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

first_successful_landing
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

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

```
%%sql select Booster_Version from SPACEXTABLE
where Landing_Outcome = 'Success (drone ship)'
and PAYLOAD_MASS__KG_ between 4000 and 6000;
```

✓ 0.0s

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

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

Total Number of Successful and Failure Mission Outcomes

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.

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

```
%%sql select Mission_Outcome, count(*) as total_number
from SPACEXTABLE group by Mission_Outcome;
```

✓ 0.0s

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

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

Boosters Carried Maximum Payload

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.

```
%%sql select Booster_Version, payload_mass__kg_
from SPACEXTABLE
where payload_mass__kg_ = (select max(payload_mass__kg_)
from SPACEXTABLE);
```

✓ 0.0s

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

2015 Launch Records

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.

```
%%sql
select strftime('%m', Date) as Month, Booster_Version,
 Launch_Site, Landing_Outcome from SPACEXTABLE
 where Landing_Outcome = 'Failure (drone ship)'
 and strftime('%Y', Date) = '2015';
```

✓ 0.0s

* sqlite:///my_data1.db
Done.

Month	Booster_Version	Launch_Site	Landing_Outcome
01	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
04	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

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

This query returns landing list that between 2010-06-04 and 2017-03-20 inclusively.

There are 4 types of landing outcome:

- Success
- Failure
- Controlled
- Uncontrolled
- Precluded
- No attempt

There were 8 successful landings, 7 failure landings, 3 controlled, 2 Uncontrolled, 1 precluded and 10 no attempt.

```
%%sql
select Landing_Outcome, count(*) as count_outcome
 from SPACEXTABLE where Date between '2010-06-04' and
  '2017-03-20' group by Landing_Outcome order by
   Count_Outcome desc;
```

✓ 0.0s

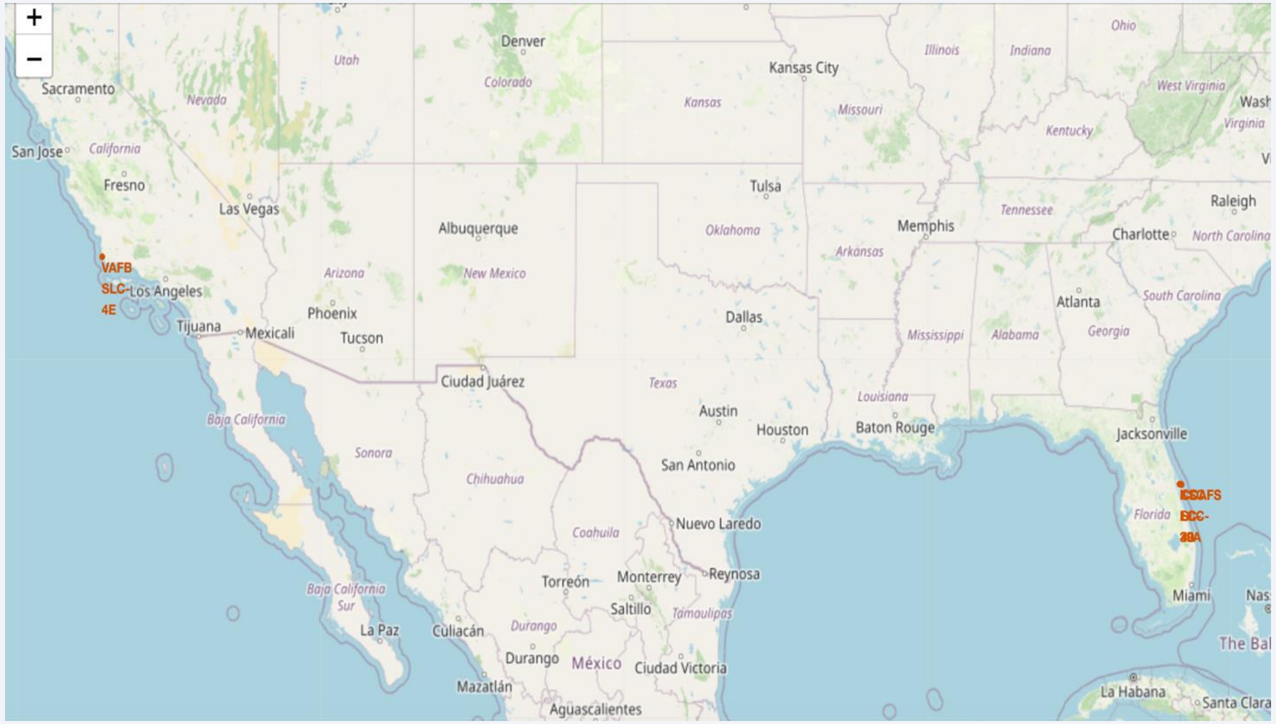
```
* sqlite:///my_data1.db
Done.
```

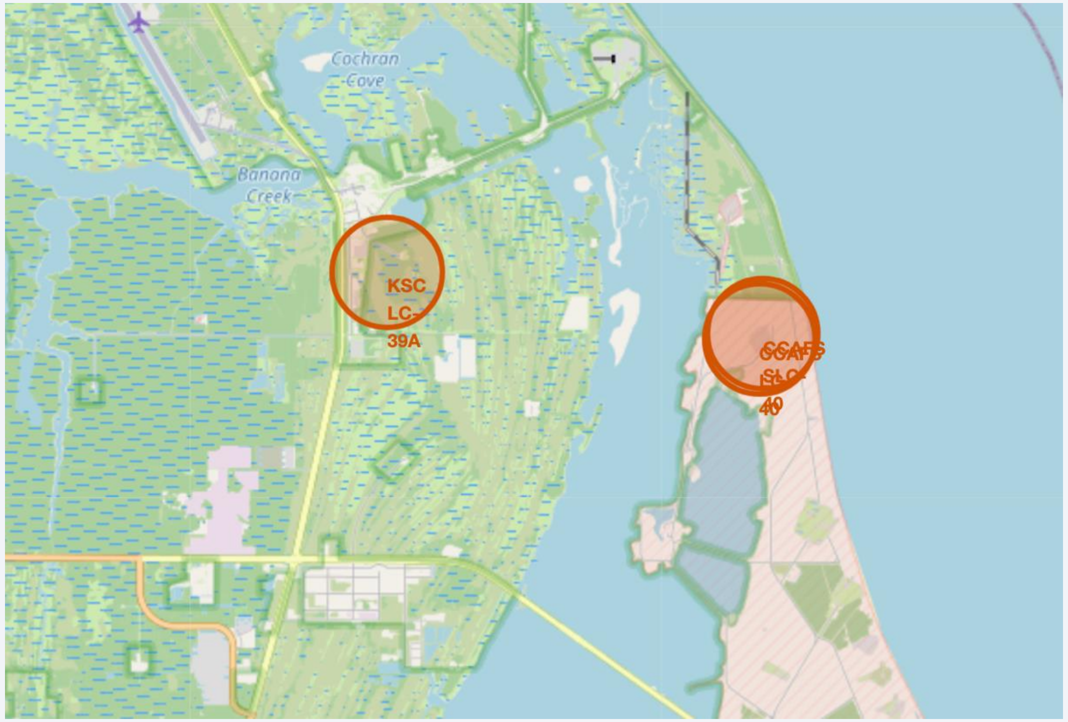
Landing_Outcome	count_outcome
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

Section 3

Launch Sites Proximities Analysis

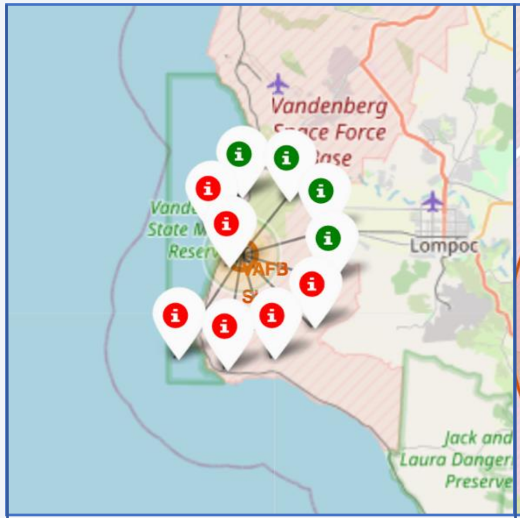
Launch Site Locations



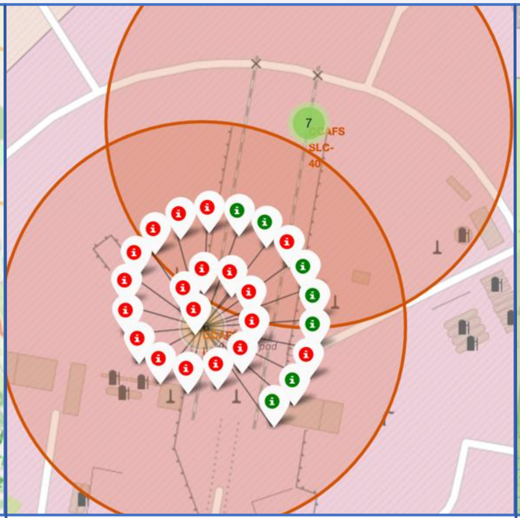


The left map shows all launch sites relative US map. The right map shows the two Florida launch sites since they are very close to each other. All launch sites are near the ocean.

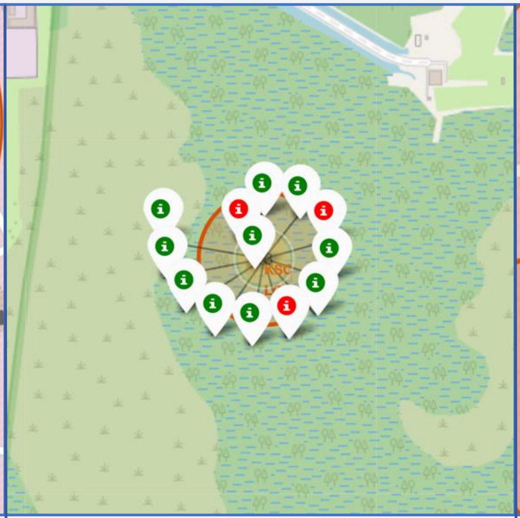
Launch Markers



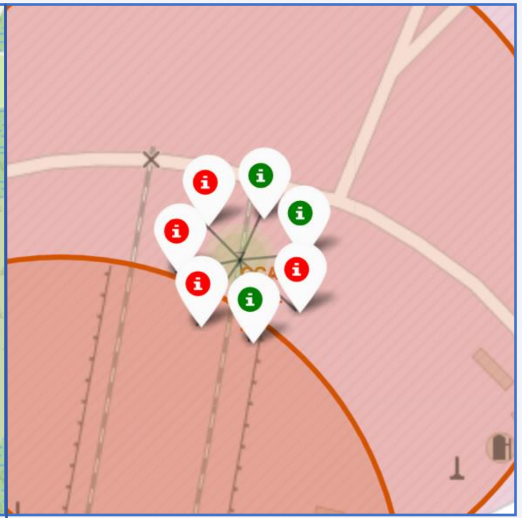
CCAFS LC-40 shows 4 successful landings and 6 unsuccessful landings as total 10landings.



CCAFS LC-40 shows 7 successful landings and 19 unsuccessful landings as total 26 landings.



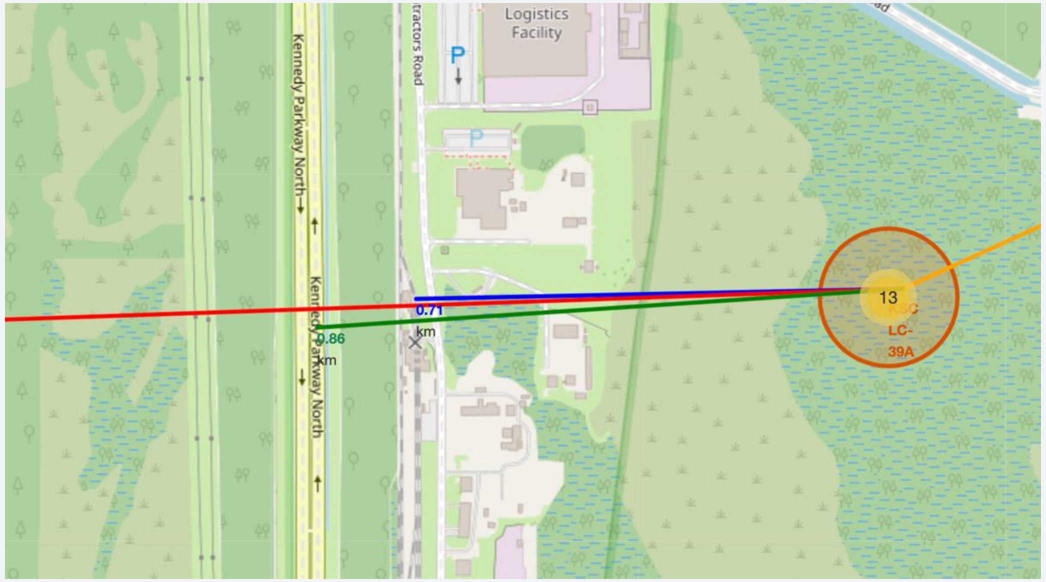
RSC LC39A shows 10 successful landings and 3 unsuccessful landings as total 13 landings.

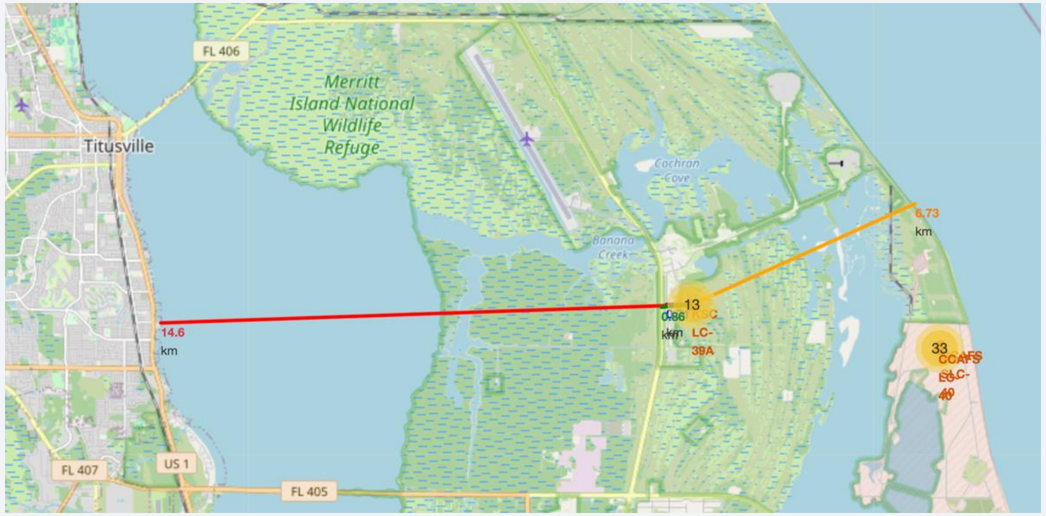


CCAFS SLC-40 shows 3 successful landings and 4 unsuccessful landings as total 7 landings.

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.

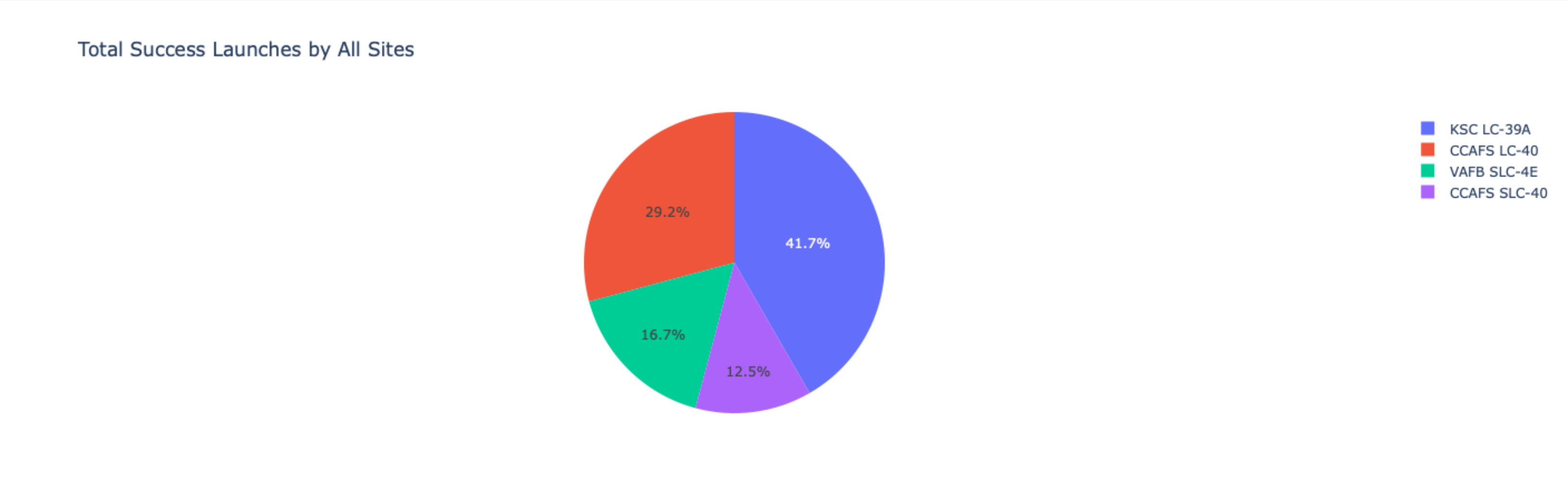




Section 4

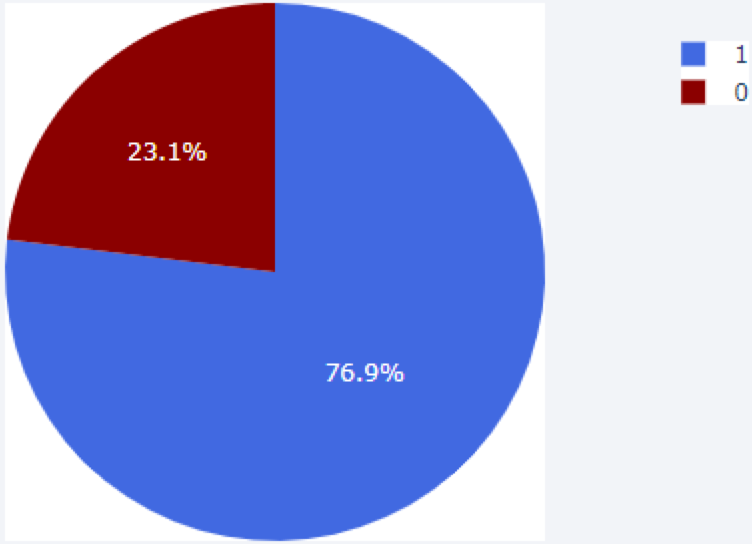
Build a Dashboard with Plotly Dash

Successful Launch Across Sites



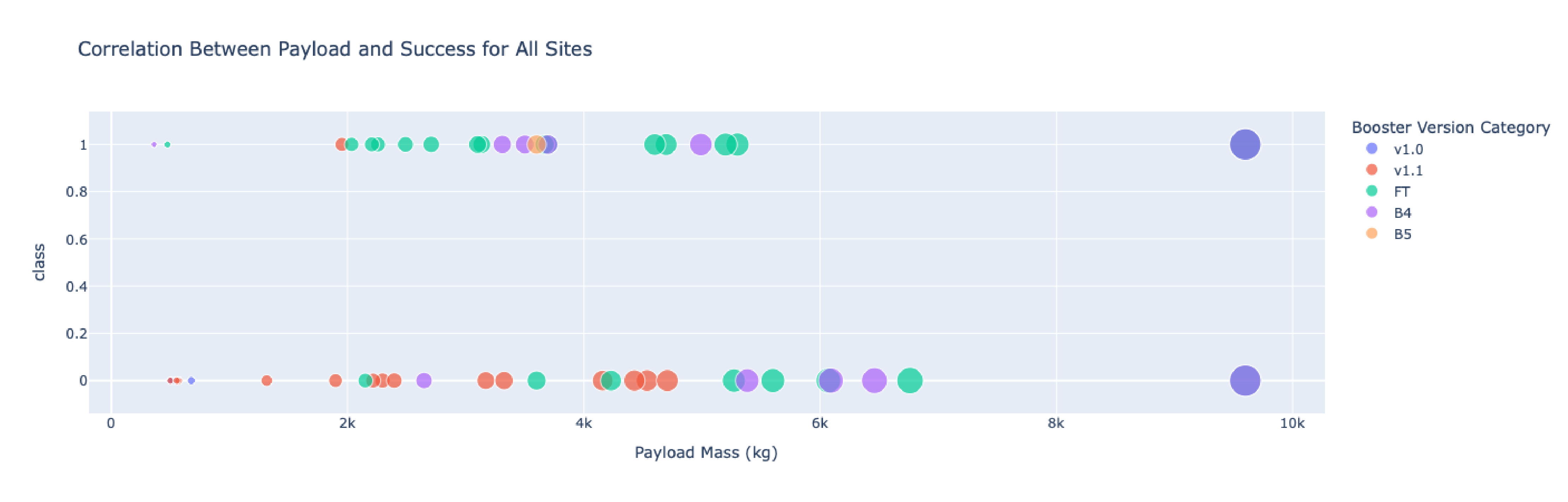
This is the distribution of successful landings across all launch sites. CCAFS LC-40 is the old name of CCAFS SLC-40 so CCAFS and KSC have the same amount of successful landings, but a majority of the successful landings where performed before the name change. VAFB has the smallest share of successful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

Highest Success Rate Launch Sute



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

Payload Mass vs Launch Outcome category



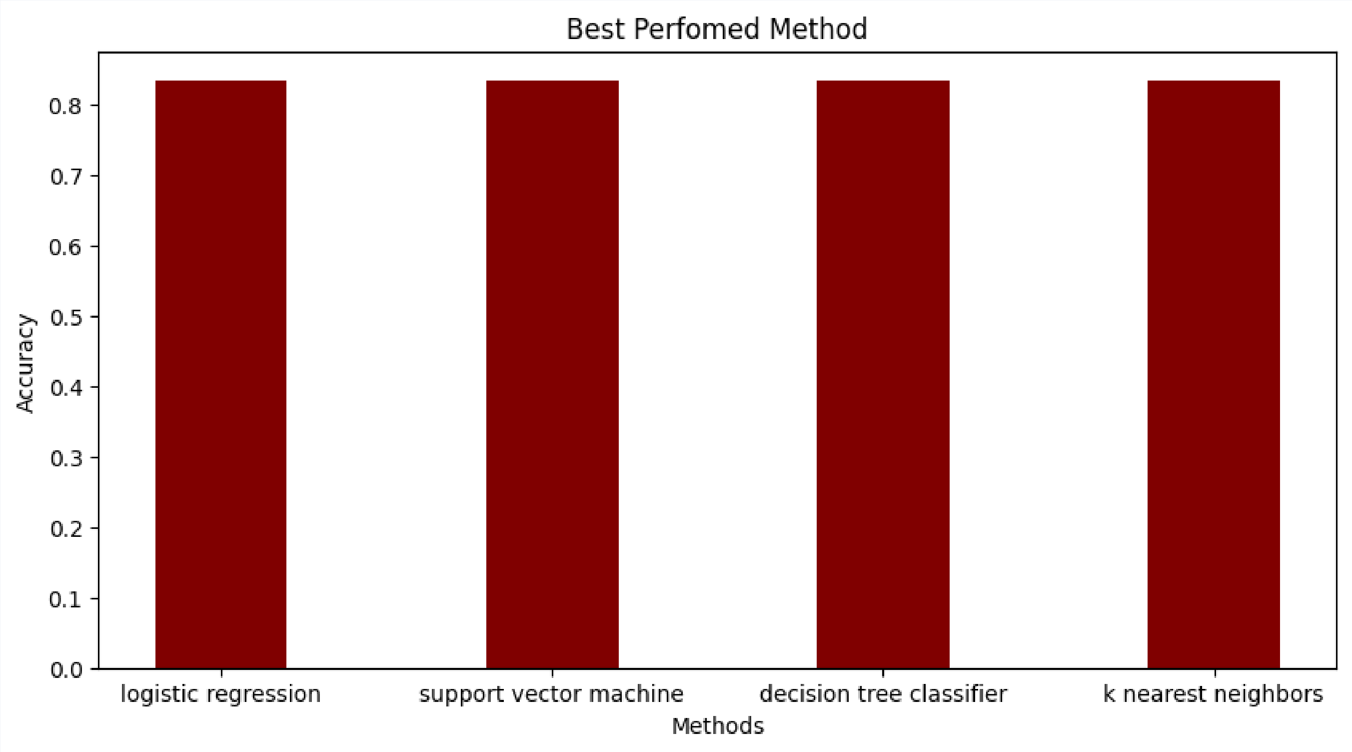
Plotly dashboard has a Payload range selector. However, this is set from 0-10000 instead of the max Payload of 15600. Class indicates 1 for successful landing and 0 for failure. Scatter plot also accounts for booster version category in color and number of launches in point size.

Section 5

Predictive Analysis (Classification)

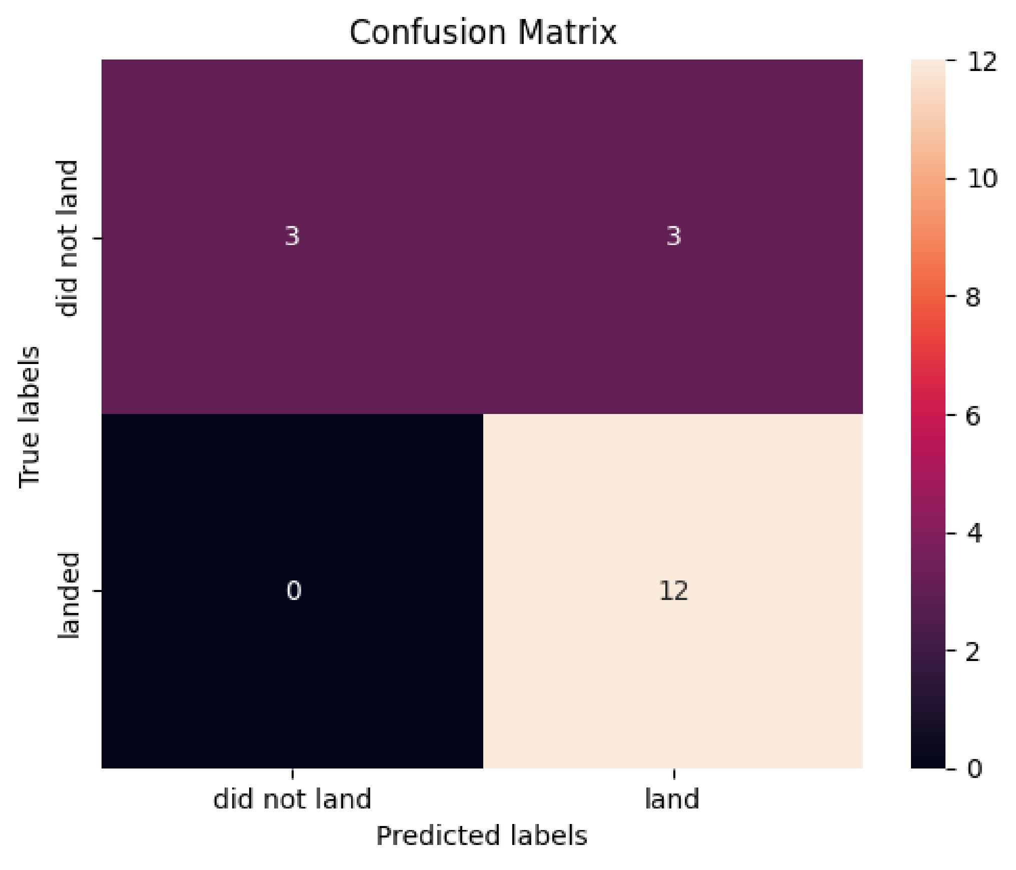
Classification Accuracy

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

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.



Conclusions

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

Appendix

GitHub repository URL:

https://github.com/Mahmoud-Mohamed-Almallah/Applied_Data_Science_IBM.git

Thank you!