IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Muhammad Zimran Khalid
30th September, 2023

Outline

Executive Summary

- **Summary of methodologies**
  - Collect data using SpaceX REST API and web scraping techniques
  - Wrangle data to create success/fail outcome variable
  - Explore data with data visualization techniques
  - Analyze the data with SQL
  - Explore launch site success rates and proximity to geographical markers
  - Visualize the launch sites with the most success and successful payload ranges
  - Build Models to predict landing outcomes using logistic regression, support vector machine (SVM), decision tree and K-nearest neighbor (KNN)
- **Summary of all results**
  - Launch success has improved over time
  - KSC LC-39A has the highest success rate among landing sites
  - Orbits ES-L1, GEO, HEO, and SSO have a 100% success rate
  - All models performed similarly on the test set. The decision tree model slightly outperformed

Introduction

- SpaceX strives to make space travel affordable for everyone.
- SpaceX can do this economically ($62 million per launch) due to its novel reuse of the first stage of its Falcon 9 rocket, as compared to its competitors, who are not able to reuse the first stage, cost upwards of $165 million each.
- By determining if the first stage will land, we can determine the price of the launch.

Explore

- How payload mass, launch site, number of flights, and orbits affect first-stage landing success
- Rate of successful landings over time
- Best predictive model for successful landing of first stage

Section 1

Methodology

Methodology

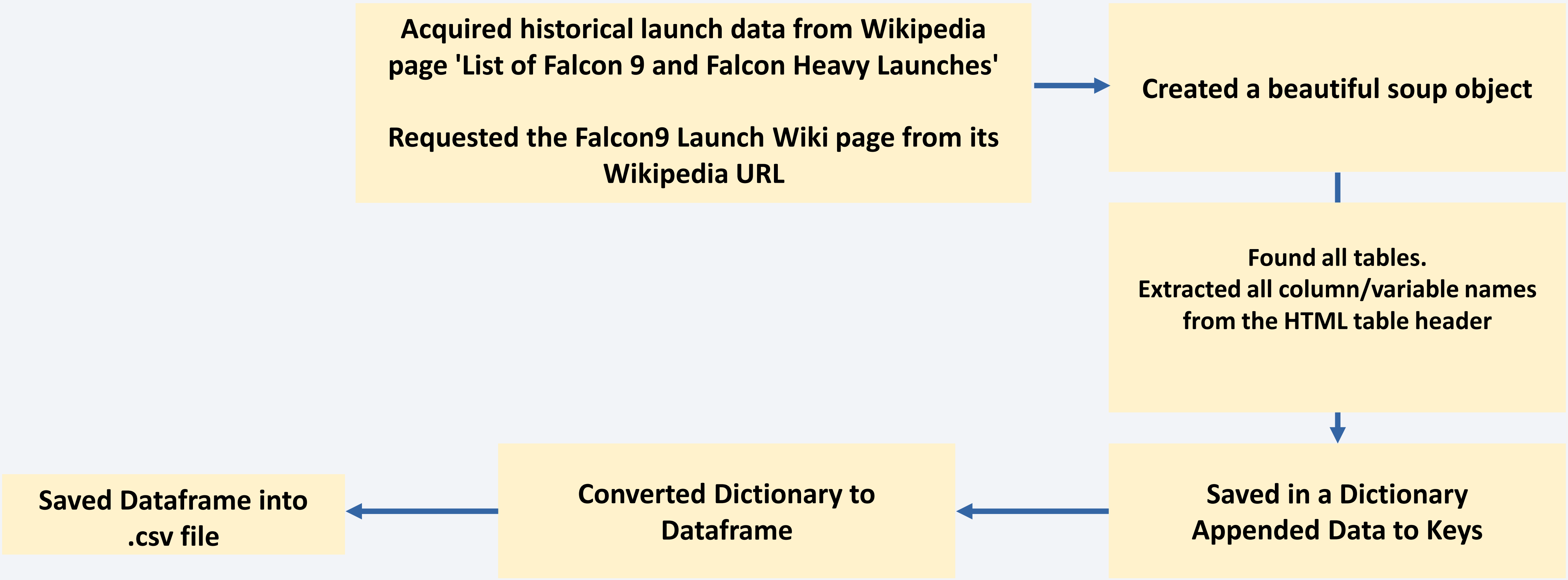
Executive Summary

- Data collection: Using SpaceX REST API and web scraping techniques
- Wrangle Data: Filtering the data, handling missing values and applying one hot encoding – to prepare the data for analysis and modeling
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
    - To predict landing outcomes using classification models. Tune and evaluate models to find best model and parameters

Data Collection

- The data collection stage is the most crucial stage in the project. Two methods were used to collect data:
  1. SpaceX API request.
  2. Web Scraping

Data Collection – Scraping



URL: [Web Scrapping]

Data Collection – SpaceX API

Requested rocket launch data from SpaceX API

Requested and parsed
SpaceX data with GET request

Decoded response as a .json file
Turned it into Pandas dataframe

Used the API with functions to get info from IDs: (BoosterVersion, Payloads, LaunchSite, Cores)

Stored in List

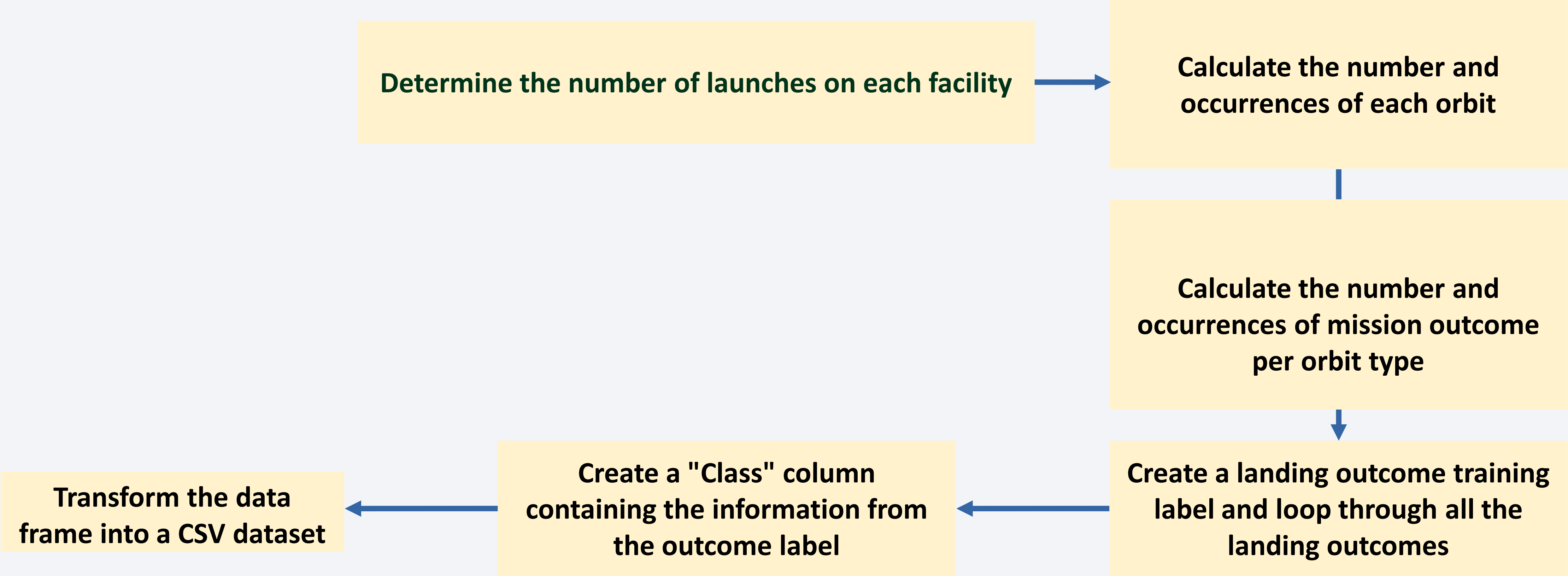
Assigned list to dictionary
Converted to Dataframe

Selected Falcon9 data only
Replaced missing payload mass value with mean

Saved Dataframe into .csv ile

URL: Data Collection

Data Wrangling



URL: [Data Wrangling]

EDA with Data Visualization

- Graphs and scatter charts with Matplotlib – Seaborn and Analysis.
- Results with Scatter charts are labeled: class 0-1 (failure/success).
- Payload mass & Flight Number
- Launch Site & Flight number
- Launch Site & Payload mass
- Orbit & Flight number
- Orbit & Payload mass
- Histogram: success rate for each orbit
- Falcon 9 & Ariane-5 launch success yearly trend.

URL: Data Visualization

EDA with SQL

- Summary of SQL queries:
- Display the names of the unique launch sites in the space mission
- Compare the payload mass with boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the total number of successful and failure mission outcomes
- Determine the dates of different landing outcomes

URL: EDA SQL

Build an Interactive Map with Folium

- Folium Markers were used to show the SpaceX launch sites and their nearest important landmarks like railways, highways, cities and coastlines.
- Polylines were used to connect the launch sites to their nearest land marks.
- Folium Circles were used to highlight circle area of launch sites.
- In order to mark the success/failed launches for each site, marker clusters were used on the map. Red represents rocket launch failures while Green represents the successes.

URL: [Folium]

Build a Dashboard with Plotly Dash

- Dropdown List with Launch Sites: Allow user to select all launch sites or a certain launch site
- Slider of Payload Mass Range: Allow user to select payload mass range
- Pie Chart Showing Successful Launches: Allow user to see successful and unsuccessful launches as a percent of the total
- Scatter Chart Showing Payload Mass vs. Success Rate by Booster Version: Allow user to see the correlation between Payload and Launch Success

URL: Dashboard

Predictive Analysis (Classification)

- **Create** NumPy array from the Class column
- **Standardize** the data with StandardScaler. Fit and transform the data.
- **Split** the data using train_test_split
- **Create** a GridSearchCV object with cv=10 for parameter optimization
- **Apply** GridSearchCV on different algorithms: logistic regression, support vector machine (SVC()), decision tree, K-Nearest Neighbor
- **Calculate** accuracy on the test data using .score() for all models
- **Assess** the confusion matrix for all models
- **Identify** the best model using Jaccard_Score, F1_Score and Accuracy

URL: [Classification]

Results

Exploratory Data Analysis

- Launch success has improved over time
- KSC LC-39A has the highest success rate among landing sites
- Orbits ES-L1, GEO, HEO and SSO have a 100% success rate

Visual Analytics

- Most launch sites are near the equator, and all are close to the coast
- Launch sites are far enough away from anything a failed launch can damage (city, highway, railway), while still close enough to bring people and material to support launch activities

Predictive Analytics

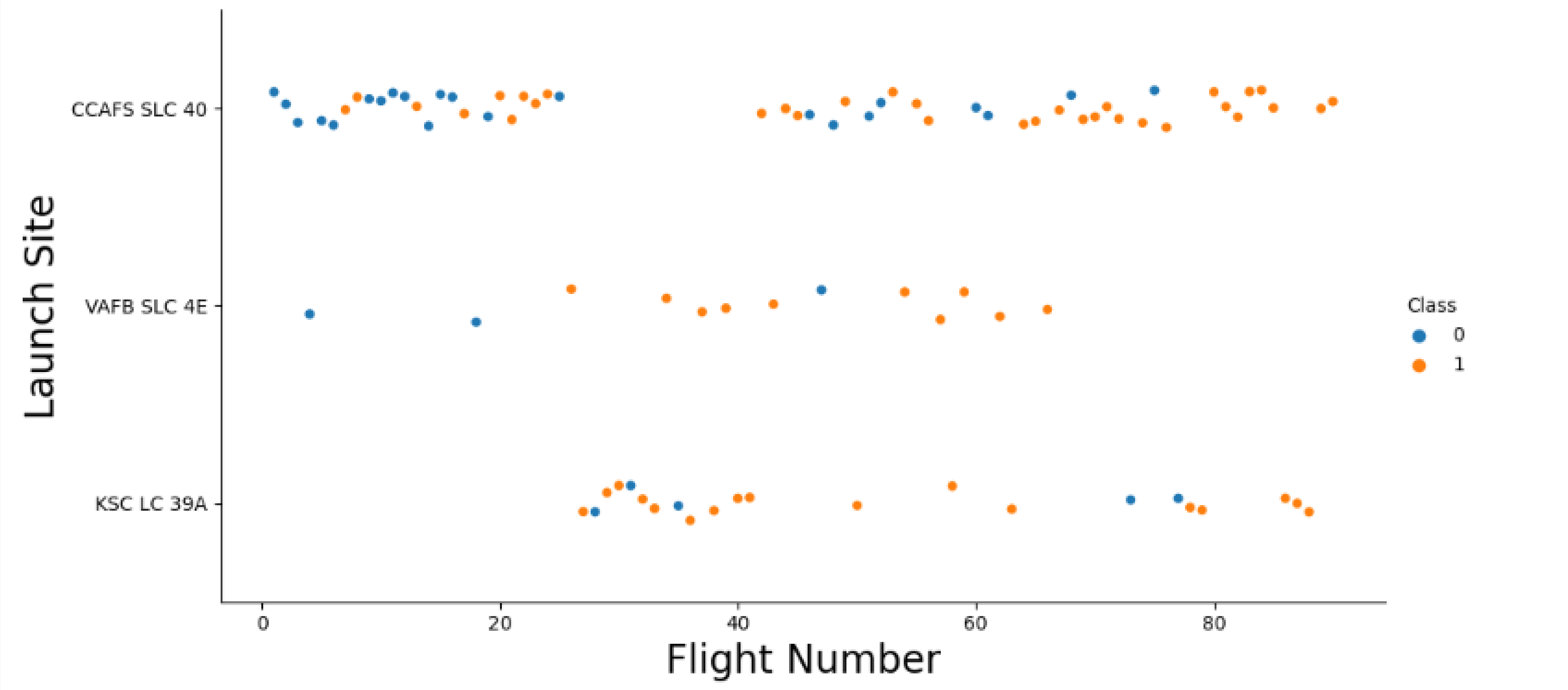
- Decision Tree model is the best predictive model for the dataset

Section 2

Insights drawn from EDA

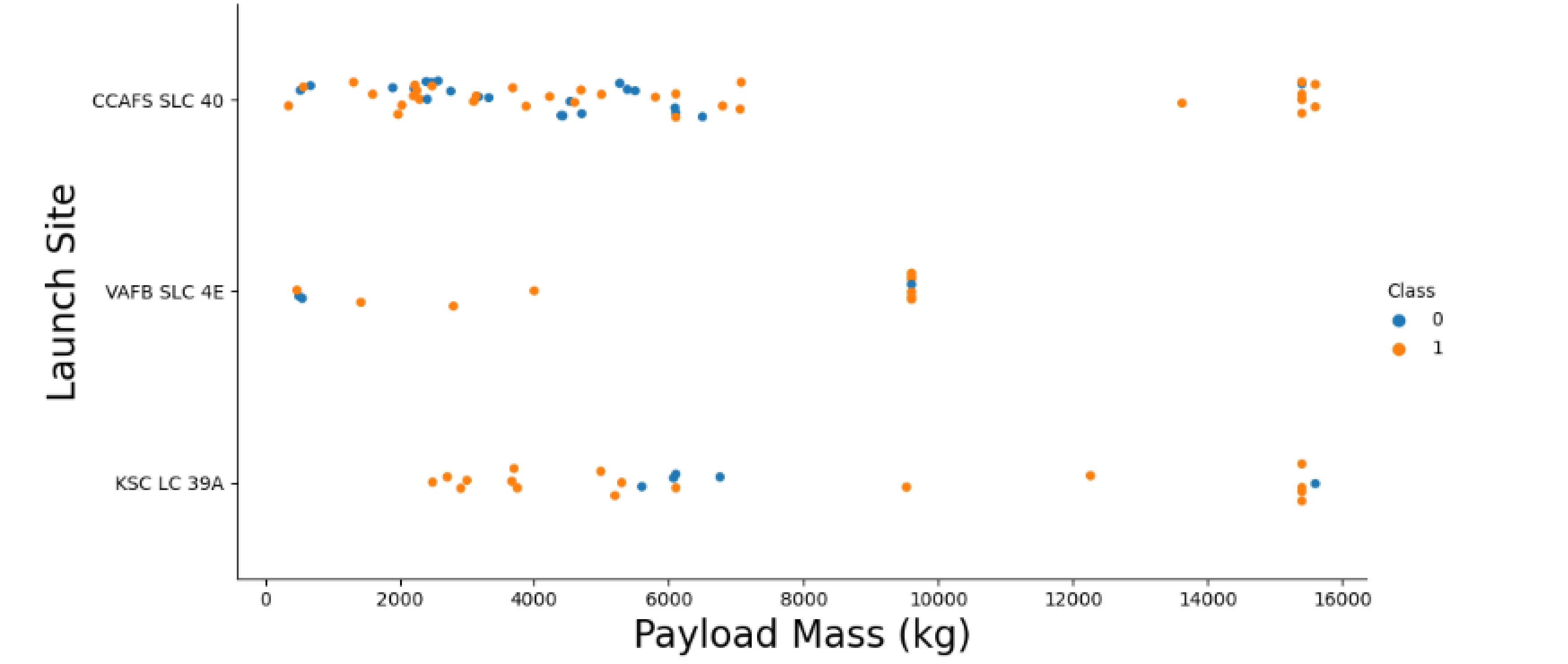
Flight Number vs. Launch Site

- Around half of launches were from CCAFS SLC 40 launch site
- VAFB SLC 4E and KSC LC 39A have higher success rates
- New launches have a higher success rate



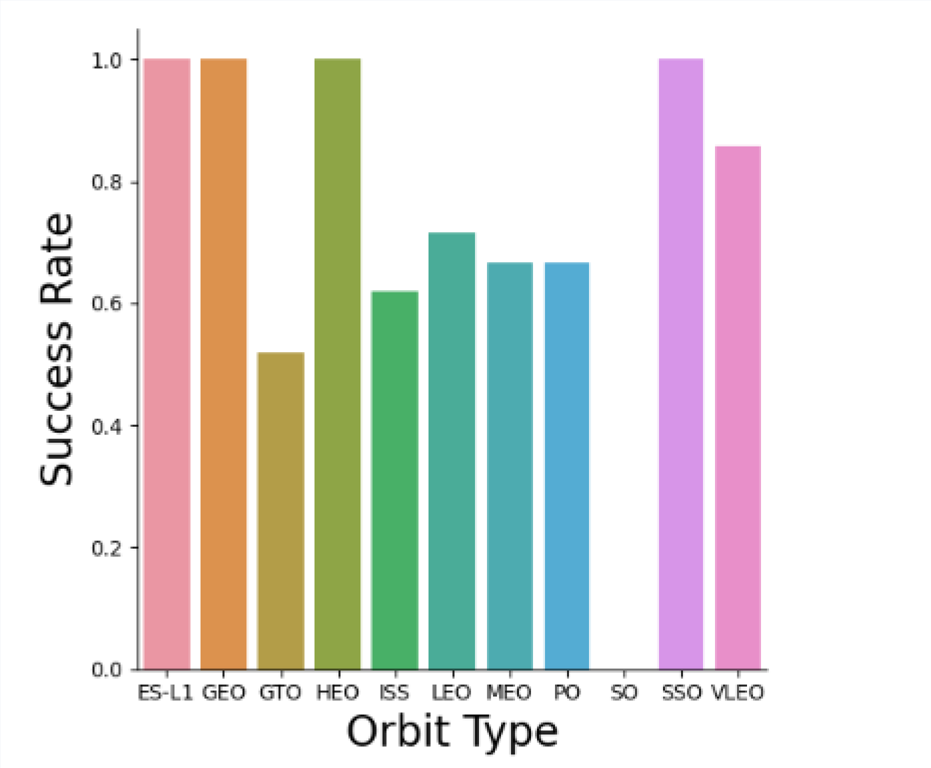
Payload vs. Launch Site

- Most launces with a payload greater than 7,000 kg were successful
- KSC LC 39A has a 100% success rate for launches less than 5,500 kg
- VAFB SKC 4E has not launched anything greater than ~10,000 kg



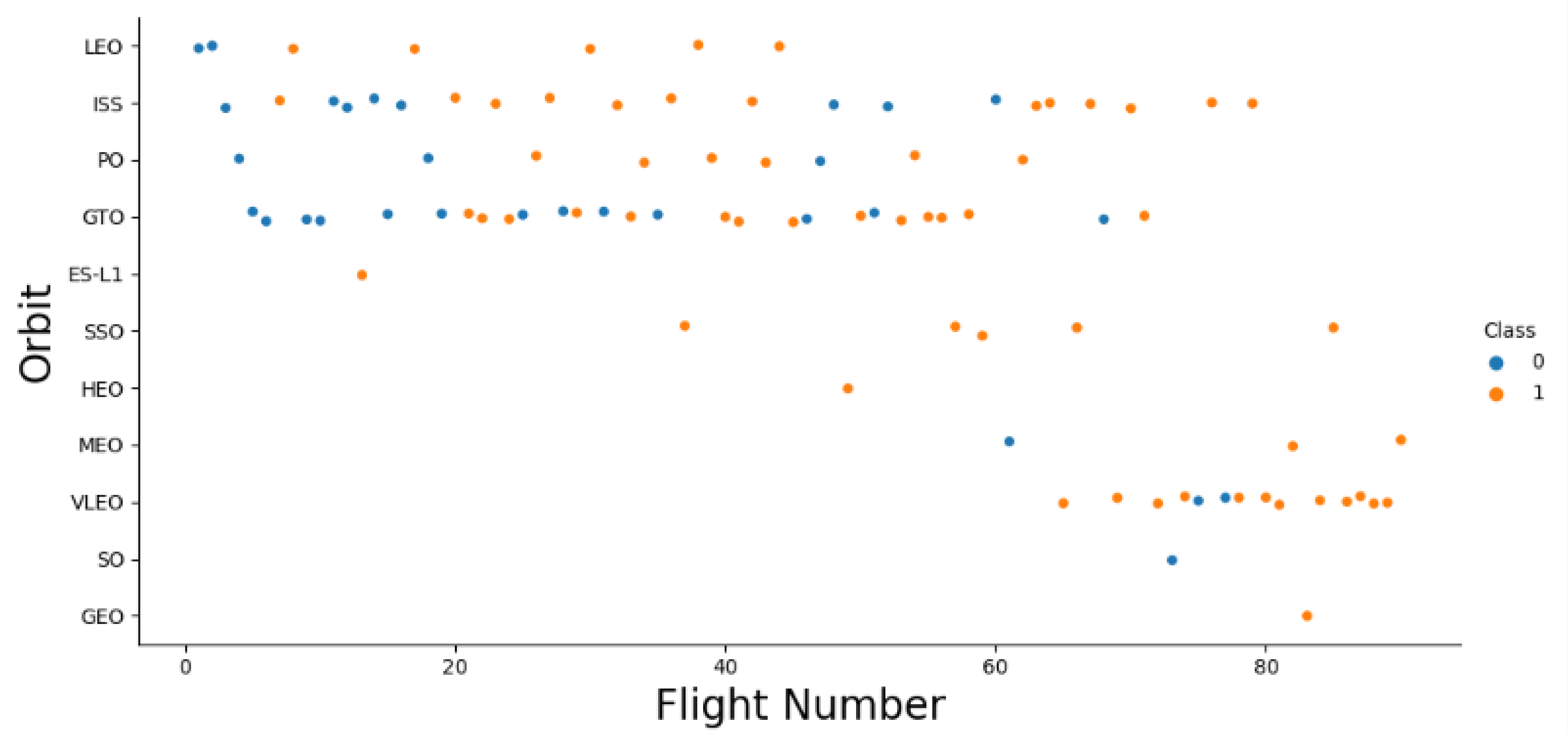
Success Rate vs. Orbit Type

- 100% Success Rate: ES-L1, GEO, HEO and SSO
- 50%-80% Success Rate: GTO, ISS, LEO, MEO, PO
- 0% Success Rate: SO



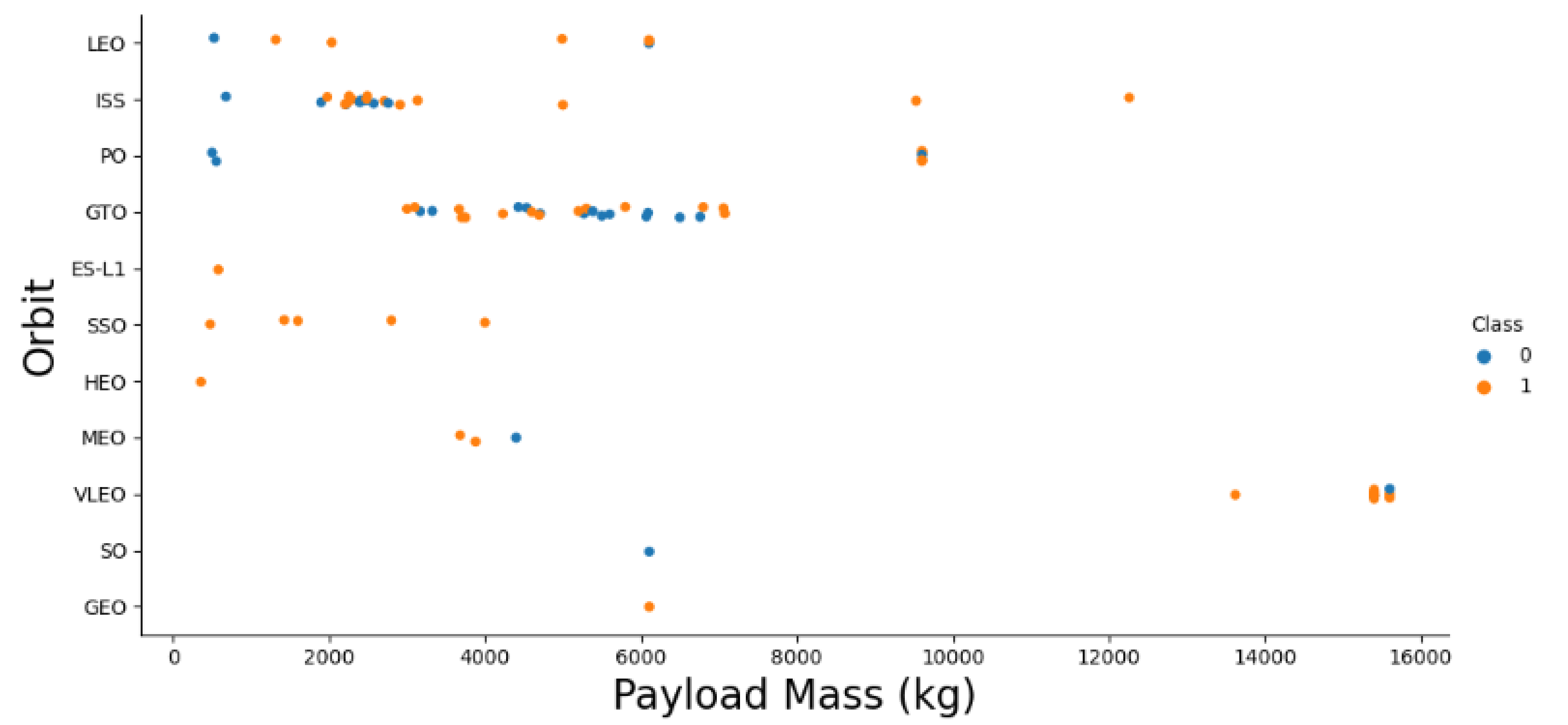
Flight Number vs. Orbit Type

- The success rate increases with the number of flights for each orbit
- This relationship is highly apparent for the LEO orbit
- The GTO orbit does not follow this trend



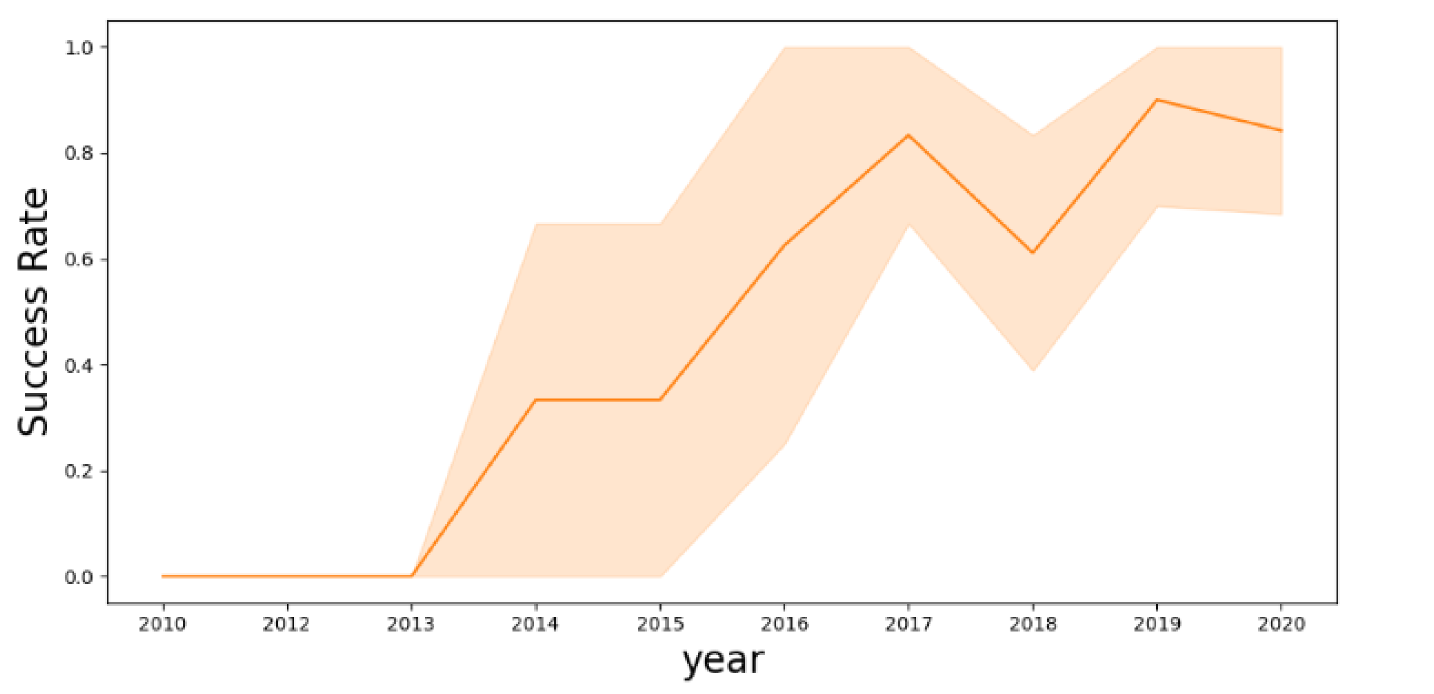
Payload vs. Orbit Type

- Heavy payloads are better with LEO, ISS and PO orbits
- GTO orbit has mixed success with heavier payloads



Launch Success Yearly Trend

- The success rate improved from 2013-2017 and 2018-2019
- The success rate decreased from 2017-2018 and from 2019-2020
- Overall, the success rate has improved since 2013



All Launch Site Names

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

In [8]:
```
sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTABLE ORDER BY 1;
```

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

Out[8]:

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

Launch Site Names Begin with 'CCA'

```
In [65]: sql SELECT * FROM SPACEXTABLE WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;
```

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

Out[65]:

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	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
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- 111,268 kg (total) carried by boosters launched by NASA

Task 3

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

In [12]:
```
sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTABLE WHERE PAYLOAD LIKE '%CRS%';
```

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

Out[12]:

TOTAL_PAYLOAD
111268

Average Payload Mass by F9 v1.1

- 2,928.4 kg (average) carried by booster version F9 v1.1

Task 4

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

In [13]:
```
sql SELECT AVG(PAYLOAD_MASS__KG_) AS AVG_PAYLOAD FROM SPACEXTABLE WHERE BOOSTER_VERSION LIKE 'F9 v1.1';
```

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

Out[13]:

AVG_PAYLOAD
2928.4

First Successful Ground Landing Date

- 1st Successful Landing in Ground Pad: 12/22/2015

Task 5

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

Hint:Use min function

In [15]:
```
sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTABLE WHERE LANDING_OUTCOME = 'Success (ground pad)'
```

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

Out[15]:

FIRST_SUCCESS_GP
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- Booster mass greater than 4,000 but less than 6,000

Task 6

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

In [27]:

```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING_OUTCOME = 'Succ
```

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

Out[27]:

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

Total Number of Successful and Failure Mission Outcomes

- 1 Failure in Flight, 99 Success, 1 Success (payload status unclear)

Task 7

List the total number of successful and failure mission outcomes

In [20]:

```
sql SELECT MISSION_OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL GROUP BY MISSION_OUTCOME ORDER BY MISSION_OUTCOME;
```

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

Out[20]:

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

Boosters Carried Maximum Payload

Task 8

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

In [30]:
```
sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL).
```

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

Out[30]:

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

2015 Launch Records

Task 9

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.

Note: SQLLite does not support monthnames. So you need to use substr(Date, 4, 2) as month to get the months and substr(Date,7,4)='2015' for year.

n [64]:

```
sql SELECT substr(DATE,6,2) as Month,BOOSTER_VERSION, LAUNCH_SITE, Landing_Outcome FROM SPACEXTBL where Landing_Outcome = 'I
```

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

ut[64]:

Month	Booster_Version	Launch_Site	Landing_Outcome
10	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

Task 10

Rank the count of landing outcomes (such as Failure (dı
20, in descending order.

In [68]:

```
sql SELECT Landing_Outcome, count(*) as count_ou
```

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

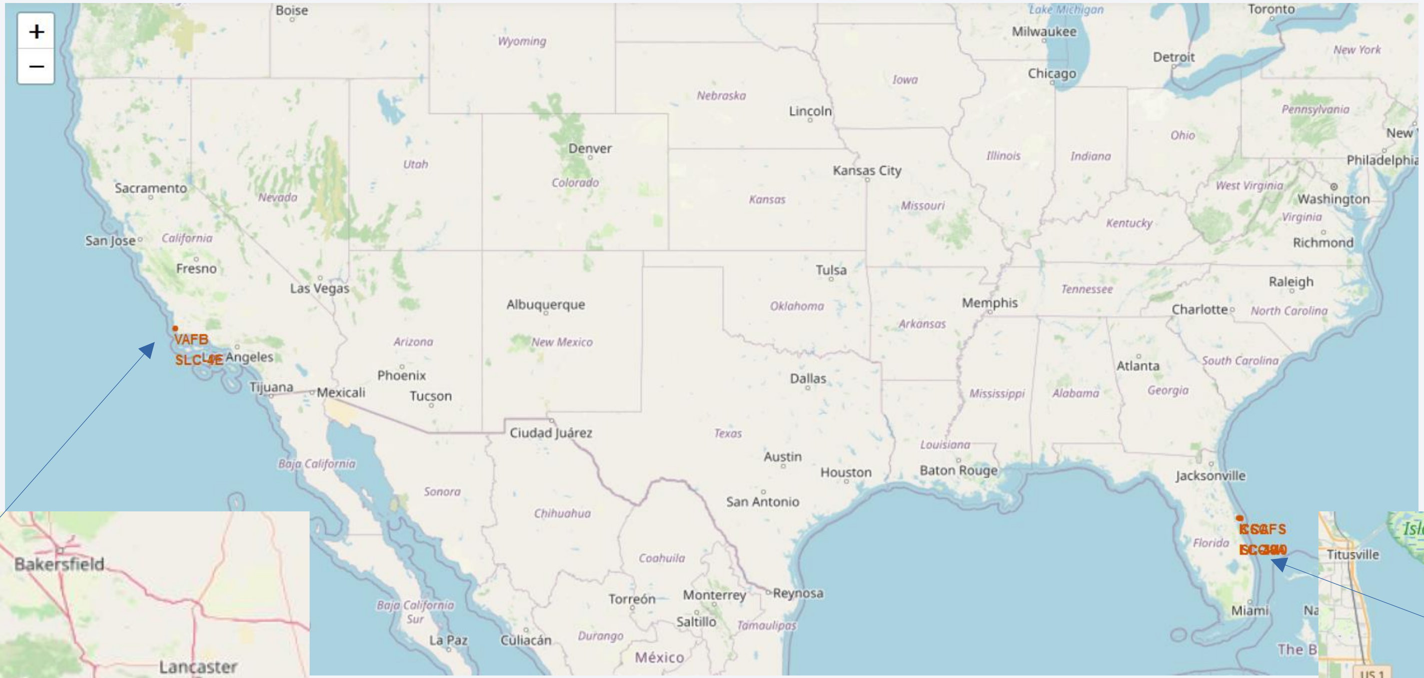
Out[68]:

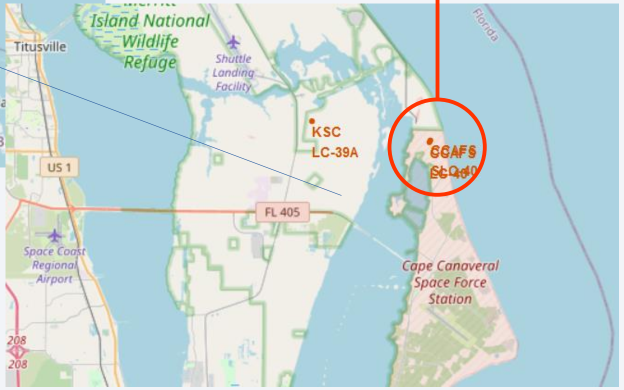
Landing_Outcome	count_outcomes
No attempt	10
Success (ground pad)	5
Success (drone ship)	5
Failure (drone ship)	5
Controlled (ocean)	3
Uncontrolled (ocean)	2
Precluded (drone ship)	1
Failure (parachute)	1

Section 3

Launch Sites Proximities Analysis

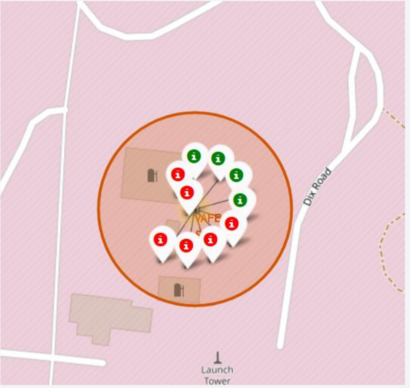
SpaceX: All launch sites



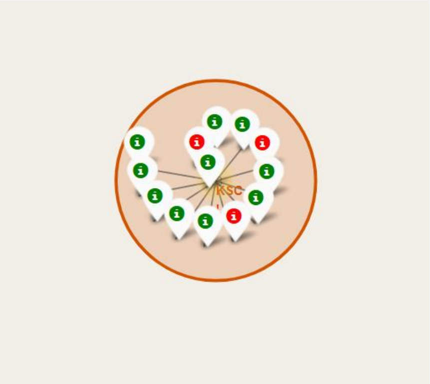


- **VAFB SLC-4E: Vandenberg Space Launch Complex 4 (CA)**
- **KSC-LC29A: Kennedy Space Center - Merritt Island (FL)**
- **CCAFS-LC40: Cape Canaveral Launch Complex 40 (FL)**
- **CCAF-SLC40: Cape Canaveral Space Launch Complex 40(FL)**

Falcon 9 Success/Failed launches for each site



Vandenberg Space Launch Complex 4 (CA)
VAFB SLC-4E



Kennedy Space Center (FL)
KSC LC 39A

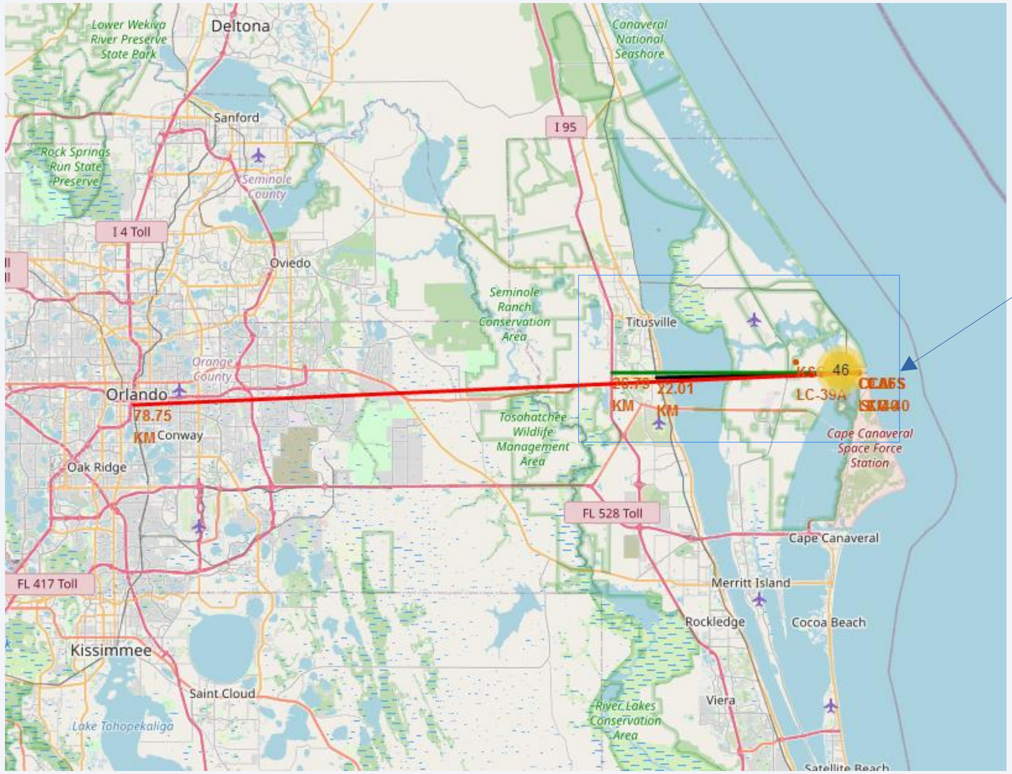
Cape Canaveral (FL)
CCAFS-LC40

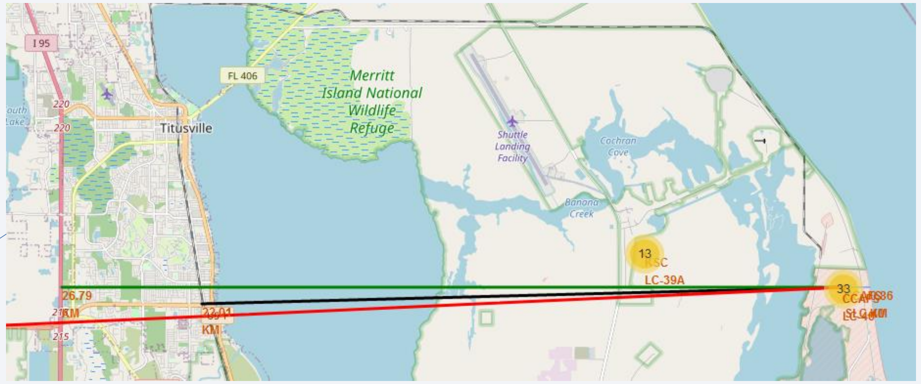
Cape Canaveral (FL)
CCAFS-SLC40

Launch Site	class	
CCAFS LC-40	0	19
	1	7
CCAFS SLC-40	0	4
	1	3
KSC LC-39A	0	3
	1	10
VAFB SLC-4E	0	6
	1	4

Table: Synthesis of launches outcomes
Class 0= failure
Class 1= success

Distances between a launch site to its proximities





Distance from CCAFS_SLC40 to:

- **Closest coast: ~900 m**
- **Florida East Coast Railway: 22.0 km**
- **Highway I 95: 26.8 km**
- **Orlando: 78.75 km**

Launch sites are close to coasts. For safety issues if launcher is lost in the early stage of the flight.

Rockets are launched:

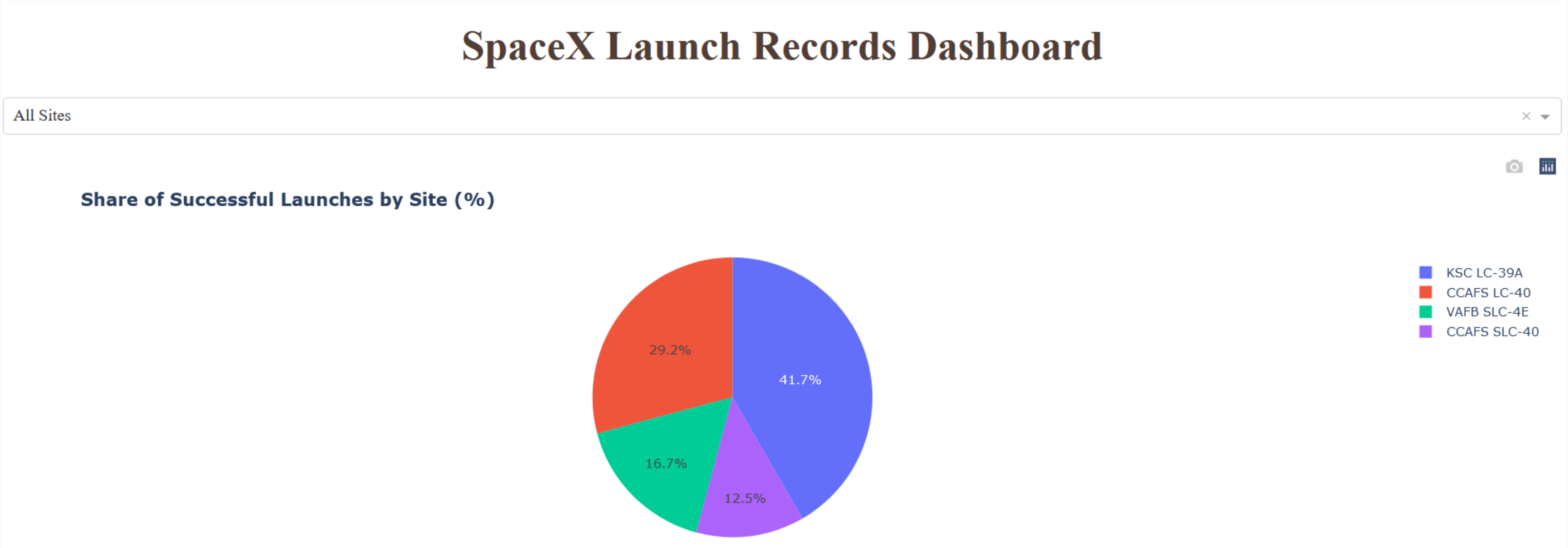
- From West to East over the ocean in Florida.
- North or South bound over the ocean in California. (Polar orbits only)

Launch sites are relatively far from populated areas for protecting population from serious incidents at lift off: explosion on the launch pad.

Section 4

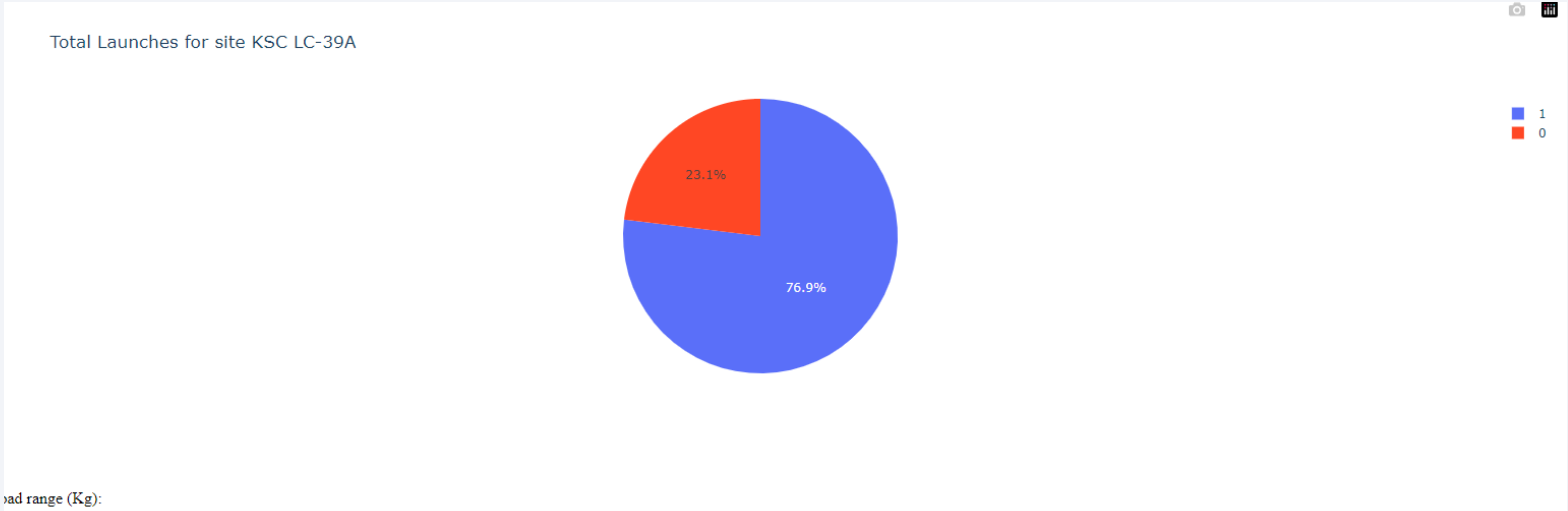
Build a Dashboard with Plotly Dash

SpaceX Falcon 9: Launch success count for all sites



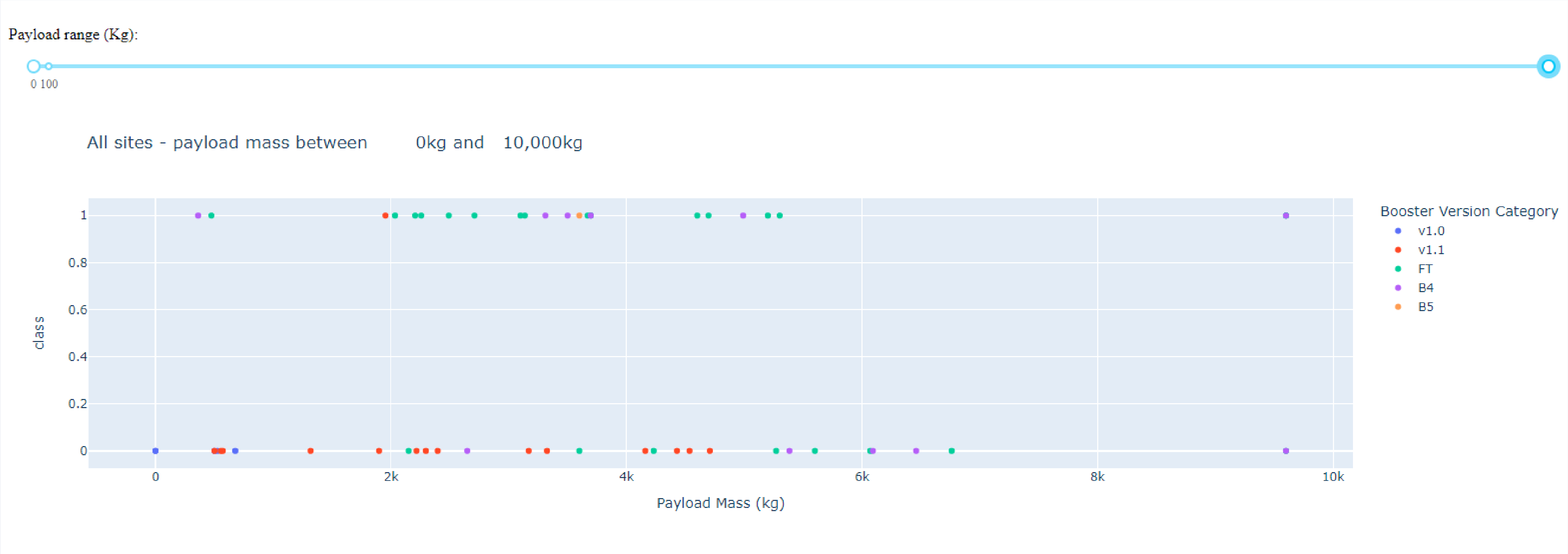
The dahsboard allows an interactive visualization and analysis of Falcon successful launches.
It completes scattered charts.
KSC LC-39A had the most successful launches from all the sites

DASHBOARD – Launch site with highest launch success ratio



KSC LC-39A achieved a 76.9% success rate while getting a 23.1% failure rate

Payload vs. Launch Outcome for all sites



We can see the success rates for low weighted payloads is higher than the heavy weighted payloads

Section 5

Predictive Analysis (Classification)

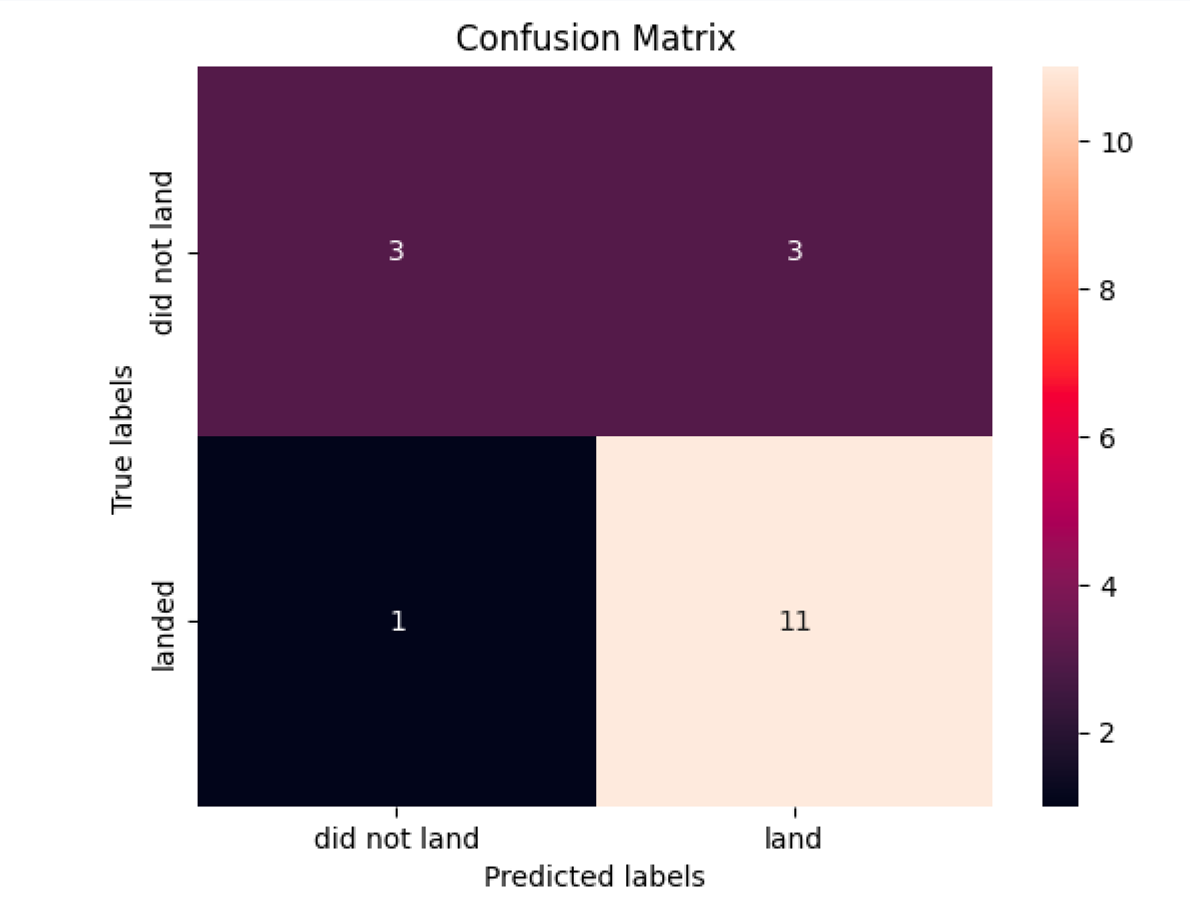
Classification Accuracy

- Tree exhibits the best accuracy: ~87%

Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.87679	0.77778
KNN	0.84821	0.83333

Confusion Matrix

- Examining the confusion matrix, we can see that Tree can distinguish between the different classes.



Conclusions

- The Tree Classifier Algorithm is the best for Machine Learning for this dataset
- Low weighted payloads perform better than the heavier payloads
- The success rates for SpaceX launches is directly proportional time in years they will eventually perfect the launches
- KSC LC-39A had the most successful launches from all the sites
- Orbit GEO,HEO,SSO,ES-L1 has the best Success Rate

Appendix

- All Source Files - https://github.com/raysengr/IBM-data-science-capstone-project.git

Thank you!