Winning Space Race with Data Science

Chukwuma John Festus

January, 2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Data on rocket launches was collected via Web APIs form Space X Webiste
- The data was cleaned and exploratory data analysis carried out with data visualization and SQL
- An interactive map and dashboard were built with Folium and Plotly Dash respectively
- Finally predictive analysis using classification models was done on the data
- The result of the analysis showed that there four unique launch site in the data namely: CCAFS SLC-40, CCAFS LC-40, VAFB SLC-4E, and KSC LC-39A with the latter having the highest success rate count of 41.7 percent
- The Tree classification model used for the model development indicated the highest accuracy with a value of 0.875

Introduction

- Project background and context
- We predicted 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; other providers cost upward of 165 million dollars each, much of the savings is because SpaceX 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.
- Problems we are looking for answers to:
- What influences if the rocket will land successfully?
- The effect each relationship with certain rocket variables will impact in determining the success rate of a successful landing.
- What conditions does SpaceX have to achieve to get the best results and ensure the best rocket success landing rate.

Section 1

Methodology

Methodology

- Data collection methodology:
  - SpaceX Rest API
  - Web Scraping form Wikipedia
- Perform data wrangling
  - Cleaning through one hot encoding, dropping irrelevant columns and creating ‘Class’ column
- Perform exploratory data analysis (EDA) using visualization and SQL
  - Generating scatter plots and bar charts to show relationships between attributes
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Building, tuning, evaluating K- Nearest Neighbors, Support Vector Machine, Logistic Regression and Decision Tree classification models

Data Collection

- To predict whether SpaceX would successfully land a rocket, datasets were collected by:
  - Gathering SpaceX launch data from the SpaceX REST API
  - Webscraping Wikipedia for Falcon 9 Launch data using BeautifulSoup
- The API gave us data about launches, including information about the rocket used, payload delivered, launch specifications, landing specifications and landing outcome
- The tables from Webscraping gave data about Falcon 9 launches such as date, time, launch sites and success or failure

Data Collection

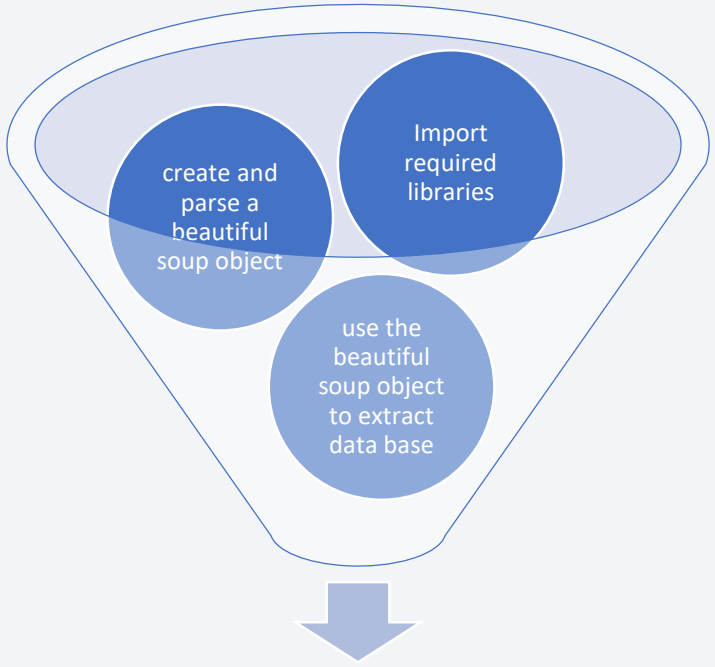
Data Collection – SpaceX API

Import required libraries and define auxiliary functions

Request and parse the rocket launch data from SpaceX API using the GET Method

https://github.com/FestusCJ/Data-Science/blob/master/Data%20Science%20Capstone%20-%20Data%20Collection%20API.ipynb

Data Collection - Scraping



https://github.com/FestusCJ/Data-Science/blob/master/Data%20Science%20Capstone%20-%20Collection%20via%20Webscraping.ipynb

Data Wrangling

- The dataframe was filtered to include only the falcon 9 launch data
- Reset the FlightNumber column using the .loc method and list, range function since some data were filtered
- Check for missing values using the.isnulland .summethods, and replace the empty rows in PayloadMass column with the mean
- Create a class column from the outcome column based on good and bad outcomes
- https://github.com/FestusCJ/Data-Science/blob/master/Data%20Science%20Capstone%20-%20Data%20Wrangling.ipynb

EDA with Data Visualization

- Categorical plot was made using Seaborn to visualize the relationship between Flight Number, Launch Site and Success rate
- Scatter plot was developed using Seaborn to visualize the following relationships:
  - Flight Number vs PayLoadMass
  - Flight Number vs Launch Site
  - PayLoadMass vs Launch Site
  - Orbit vs Flight Number
  - PayLoadMass vs Orbit
- A bar graph was used to show the relationship between Mean and Orbit
- Line plot was used to visualize the launch success yearly trend in a plot of success rate vs year
- https://github.com/FestusCJ/Data-Science/blob/master/Data%20Science%20Capstone%20-%20EDA%20with%20Visualization.ipynb

EDA with SQL

- Display the names of the unique launch sites in the mission space
- Display 5 records where launch sites begin with the string 'CCA'
- Display the total payload mass carried by the Boosters launched by NASA CRS
- Display average payload mass carried by Booster Version F9 v1.1
- Display the date where the successful landing outcome in drone ship was achieved.
- Display the names of the boosters which have success in ground pad and have payload mass greater than 4000 but less than 6000
- Display the total number of successful and failure mission outcomes
- Display the names of the booster_versions which have carried the maximum payload mass.
- Display the records which will display the month names, successful landing_outcomes in ground pad ,booster versions, launch_site for the months in year 2017
- Rank the count of successful landing_outcomes between the date 2010-06-04 and 2017-03-20 in descending order.
- https://github.com/FestusCJ/Data-Science/blob/master/Data%20Science%20Capstone%20-%20EDA%20with%20SQL.ipynb

Build an Interactive Map with Folium

- Map, markers, circles and lines objects were created using the Folium Module.
- Circle object was used to add a highlighted circle area to the Folium Map
- Markers object was used to add a text label to the Folium Map
- Lines object was used to add a connecting line from each point in the Folium Map. Lines were drawn on the map to measure distances to landmarks

- https://github.com/FestusCJ/Data-Science/blob/master/Data%20Science%20Capstone%20-%20Interactive%20Visual%20Analytics%20with%20Folium.ipynb

Build a Dashboard with Plotly Dash

- Pie Charts and Scatter Plots were created in the Dashboard
- Dropdown and Range Slider were added to the Dashboard to make it interactive
- The plots together with the interactive tools were used to obtain real time value of the visuals.

- https://github.com/FestusCJ/Data-Science/blob/master/Data%20Science%20Capstone%20Dash_interactivity.ipynb

Predictive Analysis (Classification)

- **BUILDING MODEL**
- Load our dataset into NumPy and Pandas
- Transform Data
- Split our data into training and test data sets
- Check how many test samples we have
- Decide which type of machine learning algorithms we want to use
- Set our parameters and algorithms to GridSearchCV
- Fit our datasets into the GridSearchCV objects and train our dataset.

- **EVALUATING MODEL**
- Check accuracy for each model
- Get tuned hyper parameters for each type of algorithms
- Plot Confusion Matrix

Predictive Analysis (Classification) Cont'd

- **IMPROVING MODEL**
- Feature Engineering
- Algorithm Tuning

- **FINDING THE BEST PERFORMING CLASSIFICATION MODEL**
- The model with the best accuracy score wins the best performing model
- In the notebook there is a dictionary of algorithms with scores at the bottom of the notebook.

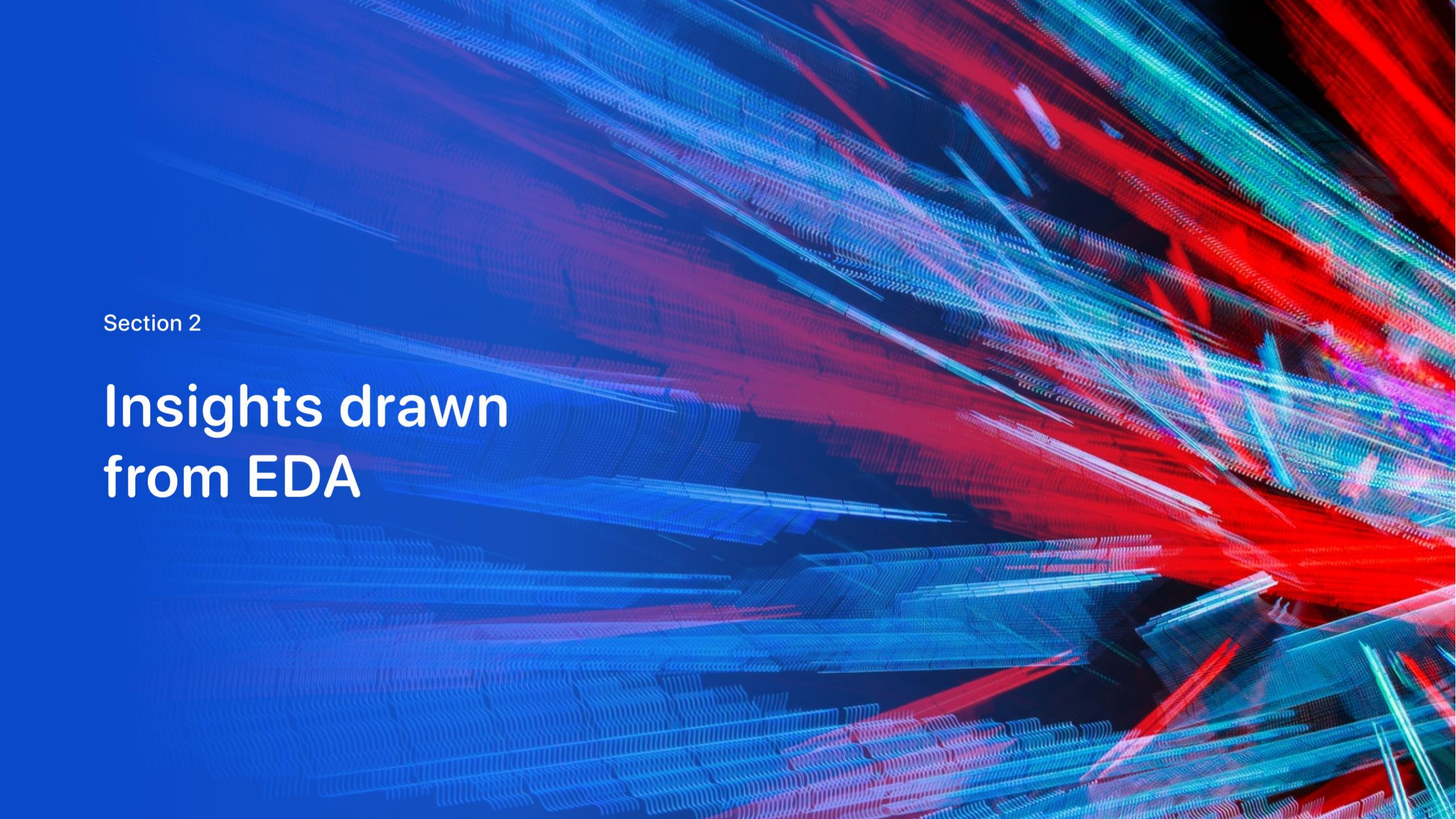
- https://github.com/FestusCJ/Data-Science/blob/master/DS%20Capstone%20-%20Machine%20Learning%20Prediction.ipynb

Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results

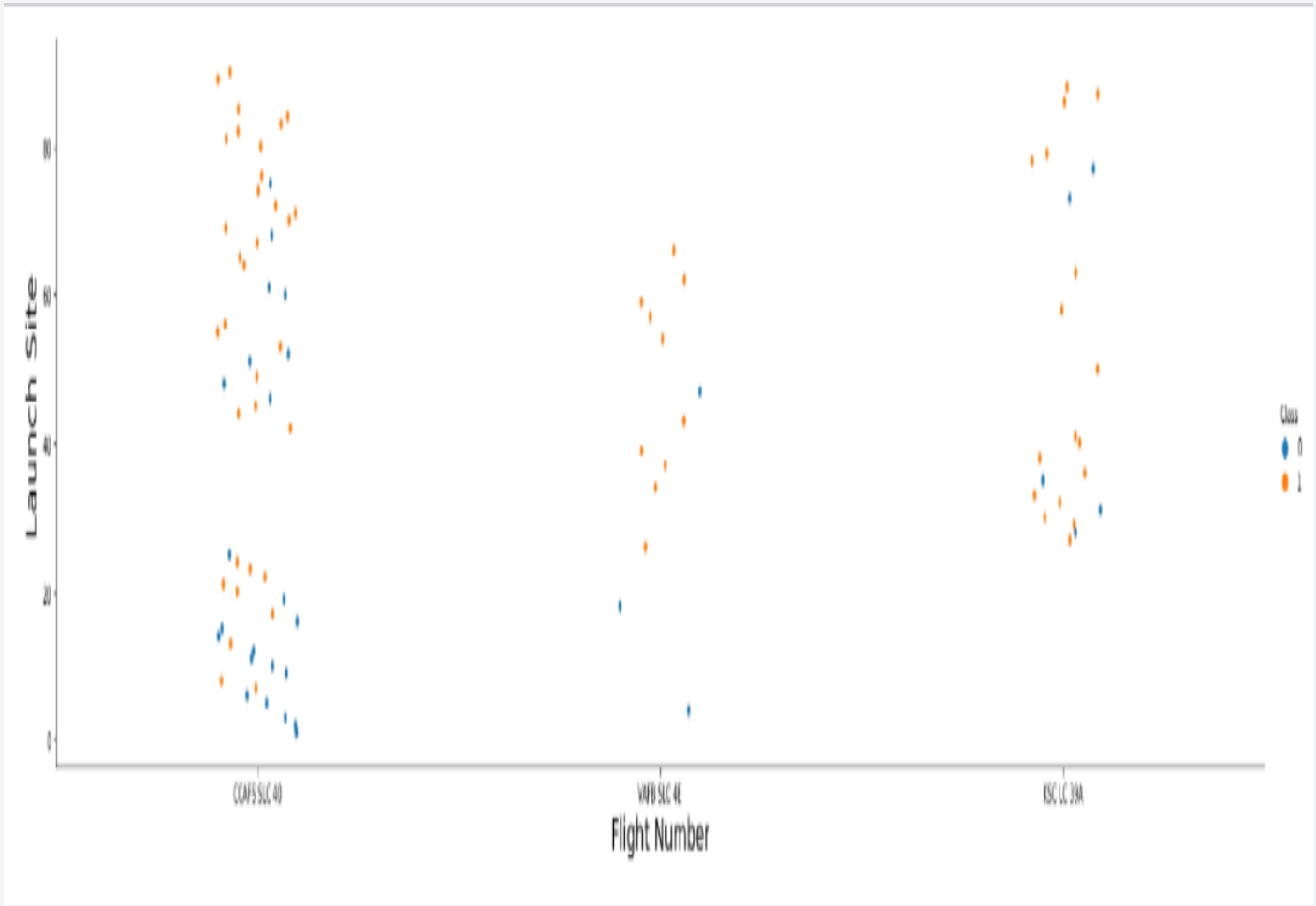
Section 2

Insights drawn from EDA



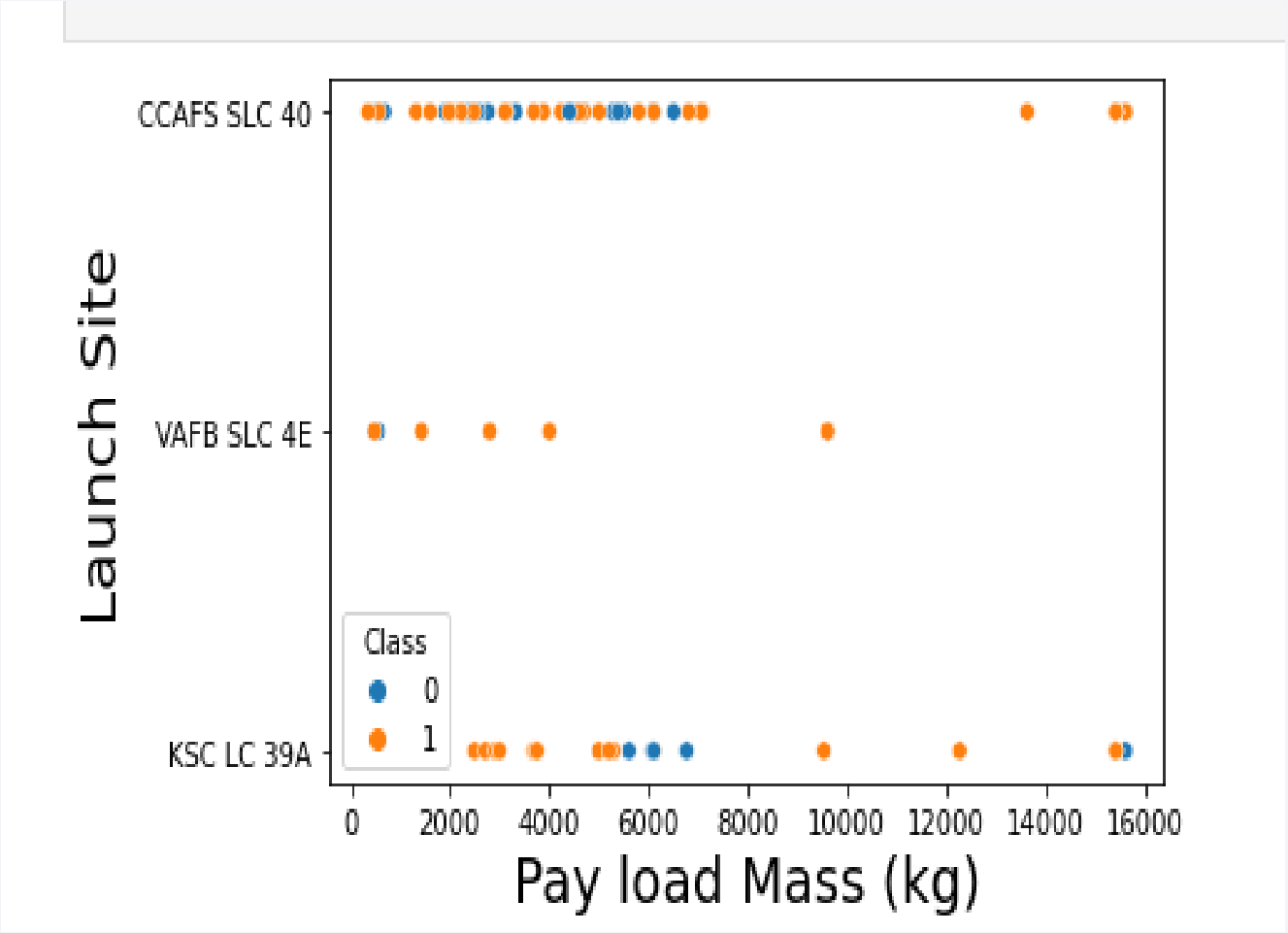
Flight Number vs. Launch Site

The scatterplot of the Flight Number versus Launch Sites indicates increase in success rate of Site VAFB as the number of flight increases. However, there tends to be no correlation between the flight number and success rate for the two sites: CCAFS and KSC



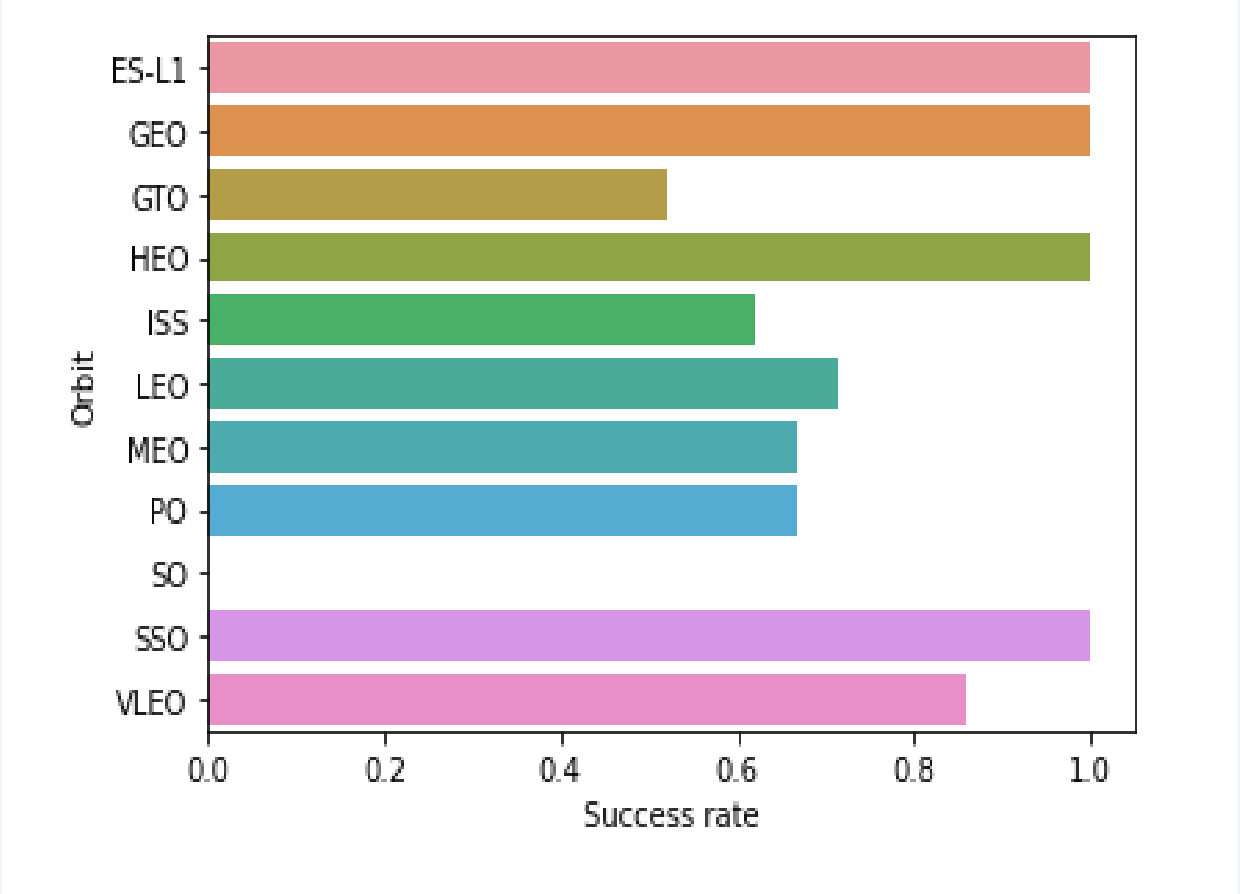
Payload vs. Launch Site

The scatter plot of Payload versus Launch site indicate a positive relationship between success rate and increase payload mass for Site VAFB.On the other hand, there is no clear relationship between success rate and an increase in payload mass.



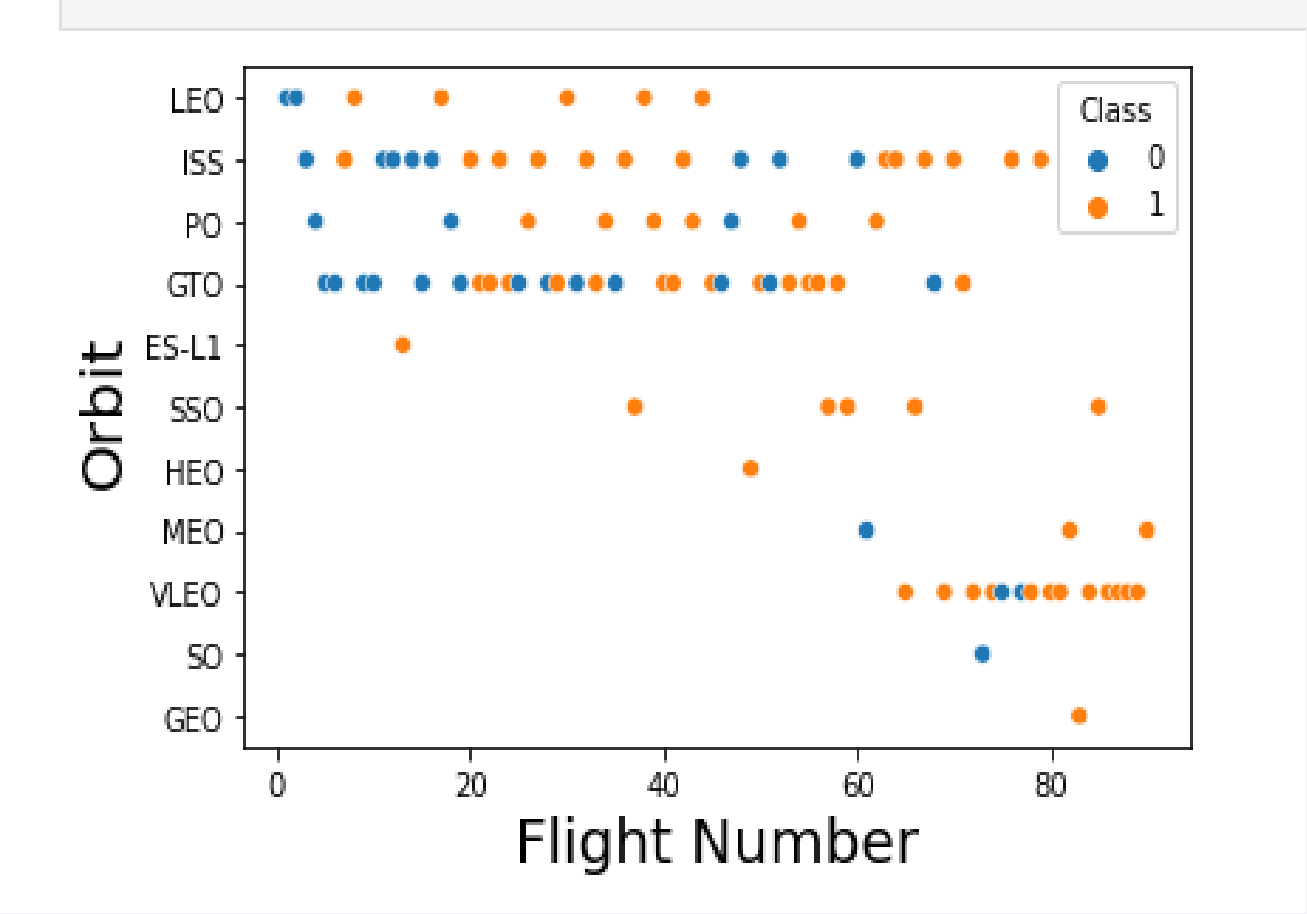
Success Rate vs. Orbit Type

While Orbits ES-L1, GEO, HEO, and SSO showed a success rate of 100%, orbit SO showed a success rate of 0%. Whereas other orbits showed varying degrees of success



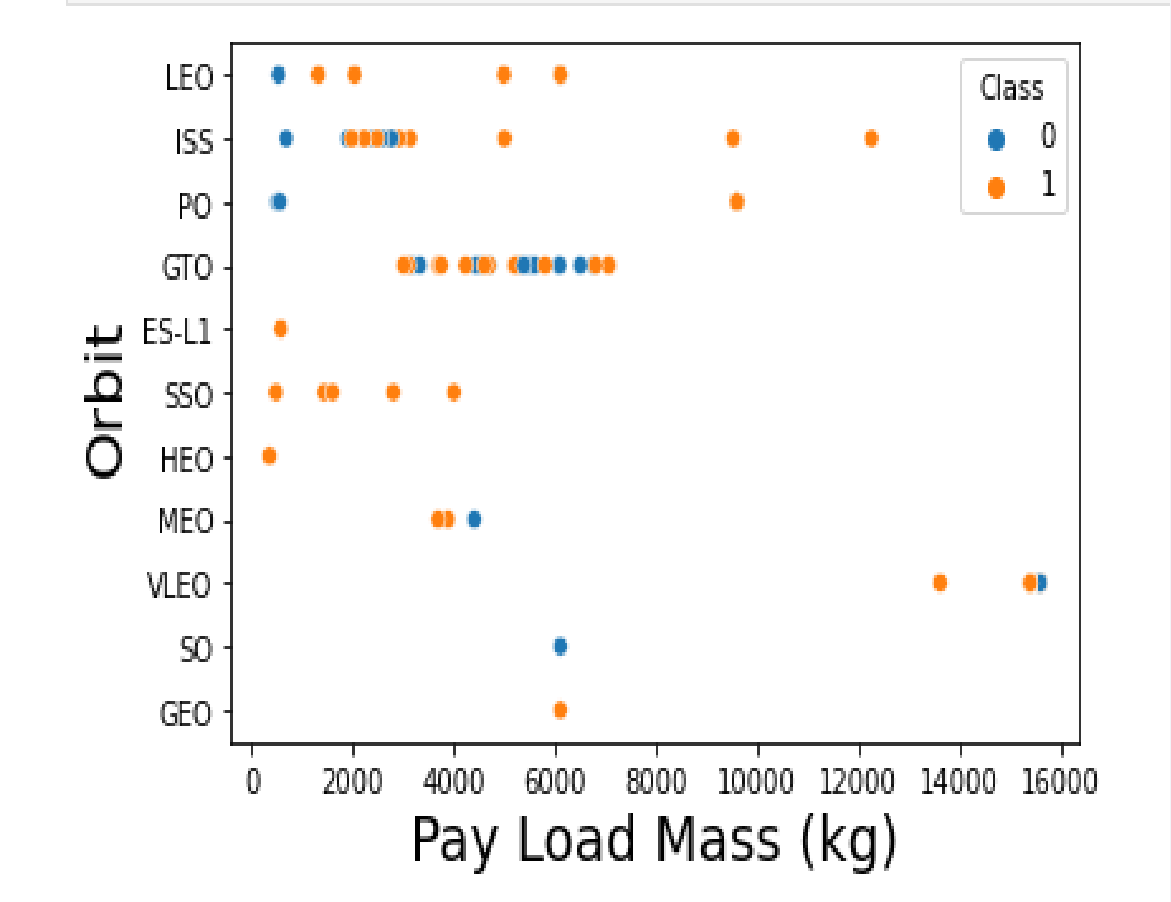
Flight Number vs. Orbit Type

Orbit LEO clearly showed a positive relationship between increase in Flight number and Success rate, likewise Orbit SSO. However there seems to be no clear correlation between increase in flight number with success rate for the other Orbits.



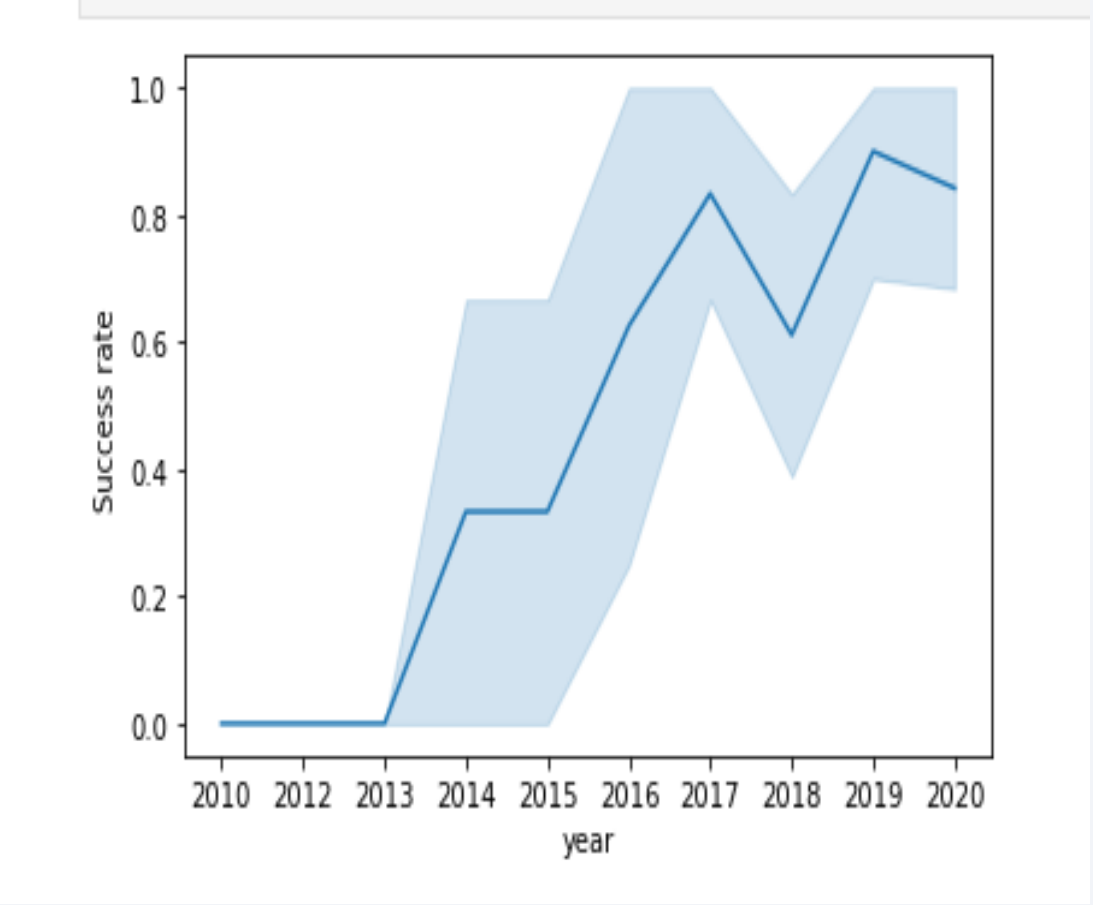
Payload vs. Orbit Type

Orbit ISS and LEO showed an increased in success rate with increase in the pay load mass. Whereas Orbit GTO did not show any clear relationship between success rate and pay load mass



Launch Success Yearly Trend

The line chart showed a positive trend with success rate meaning that as the year goes by, there tends to be more successful outcome



All Launch Site Names

- The names of the unique launch sites are displayed in the query below:

In [5]:

```
# Select relevant sub-columns: `Launch Site`, `Lat(Latitude)`, `Long(Longitude)`, `class`
spacex_df = spacex_df[['Launch Site', 'Lat', 'Long', 'class']]
launch_sites_df = spacex_df.groupby(['Launch Site'], as_index=False).first()
launch_sites_df = launch_sites_df[['Launch Site', 'Lat', 'Long']]
launch_sites_df
```

Out[5]:

	Launch Site	Lat	Long
0	CCAFS LC-40	28.562302	-80.577356
1	CCAFS SLC-40	28.563197	-80.576820
2	KSC LC-39A	28.573255	-80.646895
3	VAFB SLC-4E	34.632834	-120.610746

Launch Site Names Begin with 'CCA'

- 5 records where launch sites begin with `CCA` are displayed in the query below
- The mission outcomes were all a success and the customers were SpaceX, NASA (COTS) and NASA (CRS)

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

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

```
 * ibm_db_sa://rmy39644:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/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

Total Payload Mass

- The total payload carried by boosters from NASA is displayed in the query below:

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

In [6]:
```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD_MASS
FROM SPACEXTBL WHERE CUSTOMER = 'NASA (CRS)'
```

```
 * ibm_db_sa://rmy39644:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/BLUDB
Done.
```

Out[6]:

total_payload_mass
45596

Average Payload Mass by F9 v1.1

- The average payload mass carried by booster version F9 v1.1 is displayed in the query below:

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

In [7]:
```
%%sql

SELECT AVG(PAYLOAD_MASS__KG_) AS AV_PAYLOAD_MASS FROM SPACEXTBL
WHERE BOOSTER_VERSION = 'F9 v1.1'
```

```
 * ibm_db_sa://rmy39644:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/BLUDB
Done.
```

Out[7]:

av_payload_mass
2928

First Successful Ground Landing Date

- The date of the first successful landing outcome on ground pad is displayed in the query below:

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

Hint:Use min function

In [8]:
```
%%sql
SELECT DATE, LANDING__OUTCOME FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)' LIMIT 1
```

```
 * ibm_db_sa://rmy39644:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/BLUDB
Done.
```

Out[8]:

DATE	landing__outcome
2015-12-22	Success (ground pad)

Successful Drone Ship Landing with Payload between 4000 and 6000

- The names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000 are displayed in the query below:

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

In [9]:
```
%%sql
SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_, LANDING__OUTCOME
FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

```
 * ibm_db_sa://rmy39644:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/BLUDB
Done.
```

Out[9]:

booster_version	payload_mass__kg_	landing__outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

Total Number of Successful and Failure Mission Outcomes

- The total number of successful and failure mission outcomes are displayed in the query below:

List the total number of successful and failure mission outcomes

In [10]:
```
%%sql
SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_, LANDING__OUTCOME
FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (drone ship)' and PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

```
 * ibm_db_sa://rmy39644:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/BLUDB
Done.
```

Out[10]:

booster_version	payload_mass__kg_	landing__outcome
F9 FT B1022	4696	Success (drone ship)
F9 FT B1026	4600	Success (drone ship)
F9 FT B1021.2	5300	Success (drone ship)
F9 FT B1031.2	5200	Success (drone ship)

Boosters Carried Maximum Payload

- The names of the booster which have carried the maximum payload mass are displayed in the query below:

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

In [11]:
```
%%sql
SELECT BOOSTER_VERSION, PAYLOAD_MASS__KG_ FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

```
 * ibm_db_sa://rmy39644:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/BLUDB
Done.
```

Out[11]:

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

- The failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015 are displayed in the query below:

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

In [12]:
```
%%sql
SELECT YEAR(DATE) AS YEAR, BOOSTER_VERSION, LANDING__OUTCOME, LAUNCH_SITE FROM SPACEXTBL
WHERE EXTRACT(YEAR FROM DATE) = 2015 AND LANDING__OUTCOME = 'Failure (drone ship)'
```

```
 * ibm_db_sa://rmy39644:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/BLUDB
Done.
```

Out[12]:

YEAR	booster_version	landing__outcome	launch_site
2015	F9 v1.1 B1012	Failure (drone ship)	CCAFS LC-40
2015	F9 v1.1 B1015	Failure (drone ship)	CCAFS LC-40

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

- The count rank of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order are displayed in the query below:

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

In [13]:
```
%%sql
SELECT LANDING__OUTCOME, COUNT(LANDING__OUTCOME) AS COUNT
FROM SPACEXTBL GROUP BY LANDING__OUTCOME ORDER BY COUNT DESC
```

```
 * ibm_db_sa://rmy39644:***@2f3279a5-73d1-4859-88f0-a6c3e6b4b907.c3n41cmd0nqnrk39u98g.databases.appdomain.cloud:30756/BLUDB
Done.
```

Out[13]:

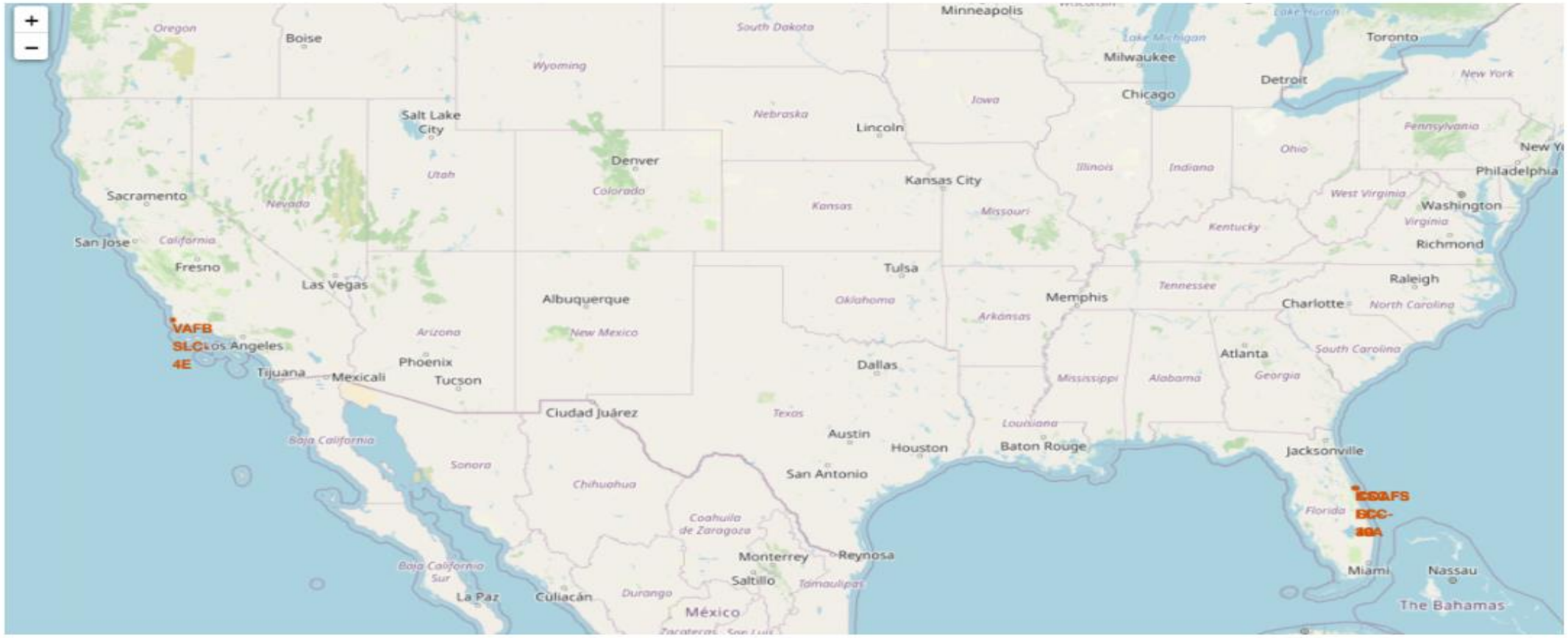
landing__outcome	COUNT
Success	38
No attempt	22
Success (drone ship)	14
Success (ground pad)	9
Controlled (ocean)	5
Failure (drone ship)	5
Failure	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1

Section 4

Launch Sites Proximities Analysis

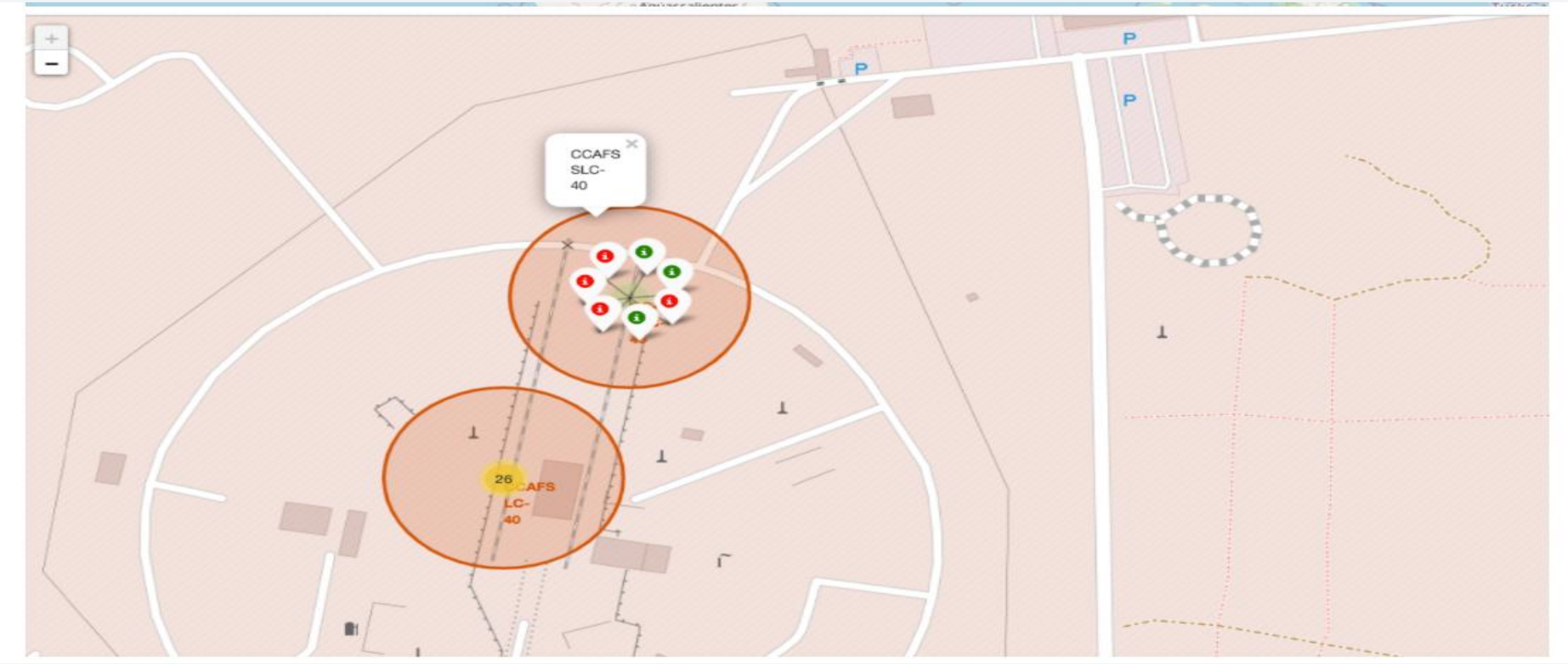
Map locations of launch sites

- VAFB is located at the coast of Los Angeles whereas the other three sites: CCAFSLC-40, CCAFSSLC-40, and KSCLC-39A are located in the coast of Miami.



Color labeled Markers

- The green markers indicate successful outcomes while the red markers indicate failed outcomes.

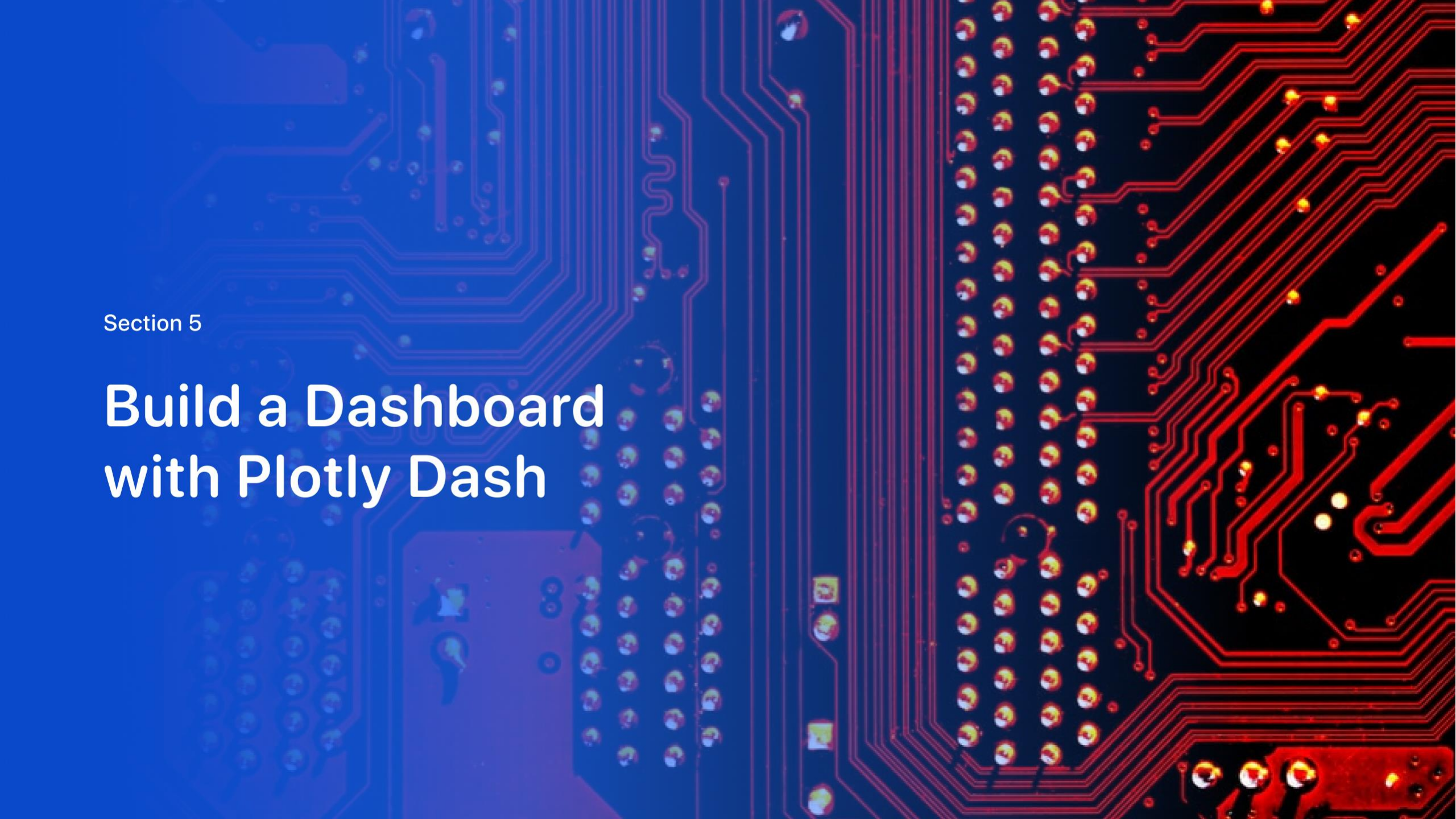


Proximities of landmarks to launch sites

- The distance from the CCAFS SLC-40 to the coastline, highway, railway, and city of Melbourne is 1.06, 1.14, 1.31, and 51.24Km respectively

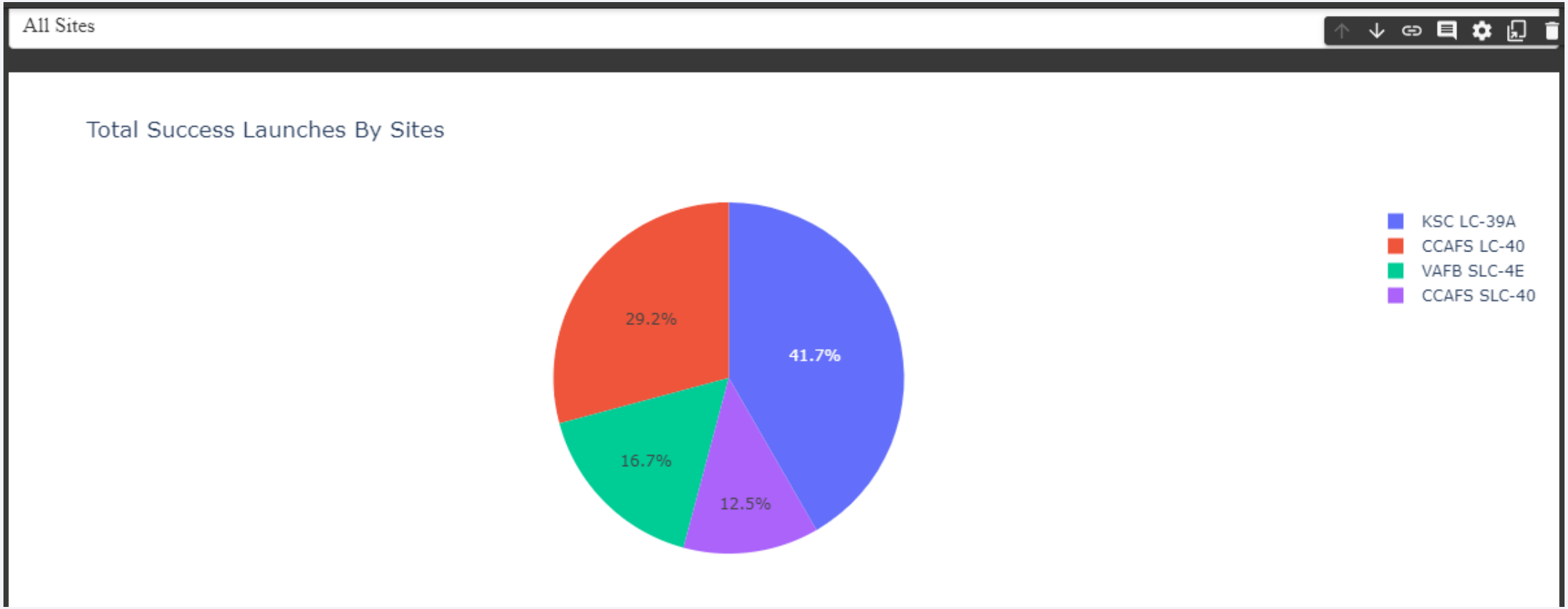
Section 5

Build a Dashboard with Plotly Dash



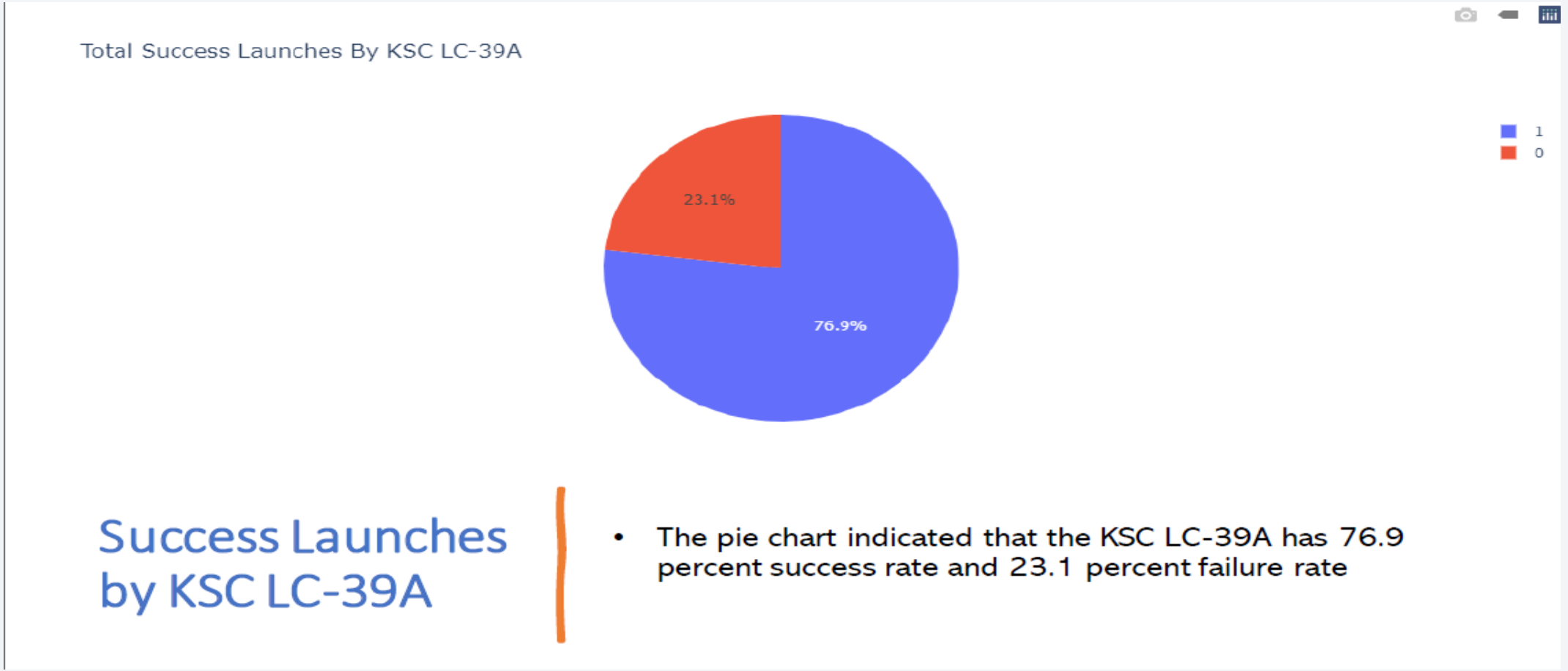
Pie Chart of Launch success count for all sites

- The pie chart indicated that KSC LC-39A site has the highest success count with 41.7% and the CCAFS SLC-40 has the lowest success count with 12.5%



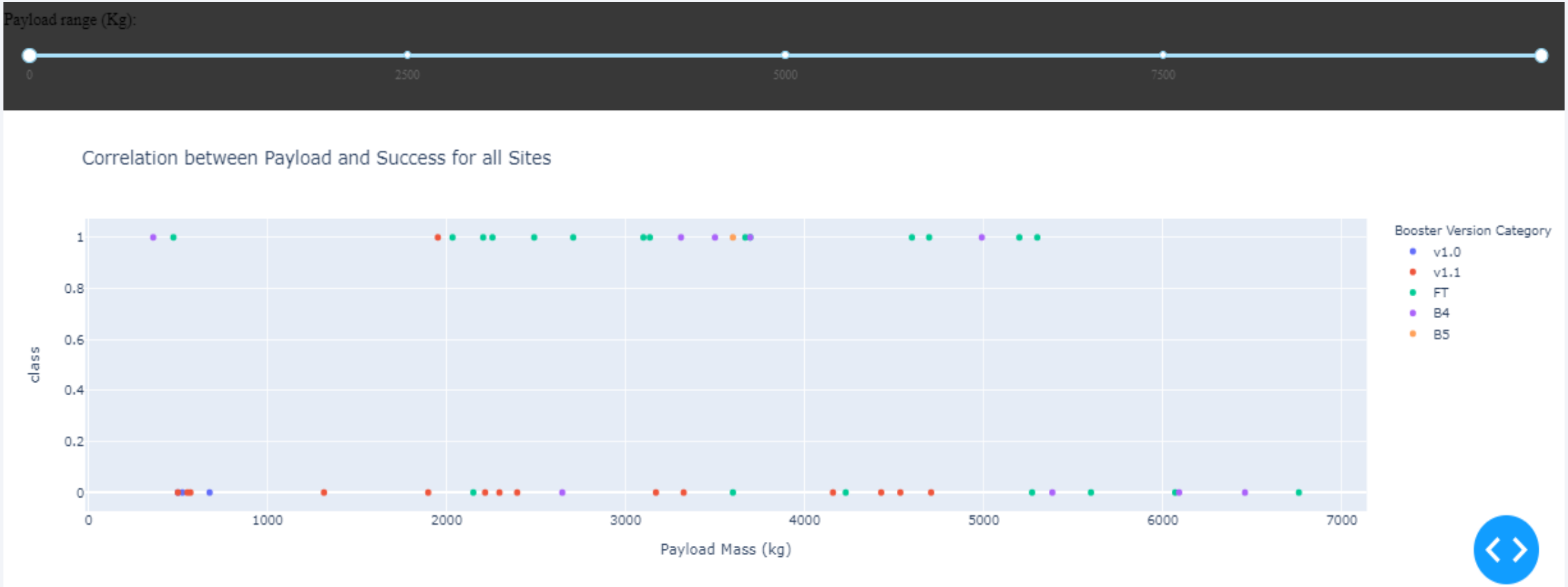
Launch site with highest launch success ratio

- The pie chart indicated that the KSC LC-39A has 76.9% success rate and 23.1% failure rate



Plot of Payload vs Launch Outcome

- The scatter plot for all sites showed the various booster version categories and payload range between 2500 and 7500. This indicated that the FT Version had the highest success rate

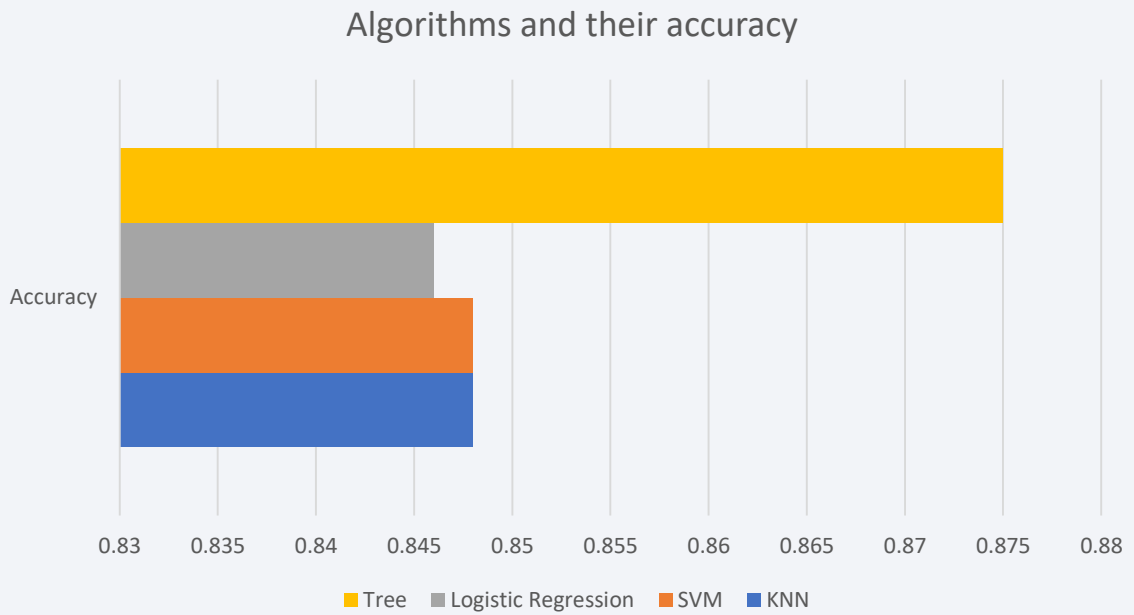


Section 6

Predictive Analysis (Classification)

Classification Accuracy

- After analysis using relevant queries, the decision tree came out best with an accuracy of 87.5%



Find the method performs best:

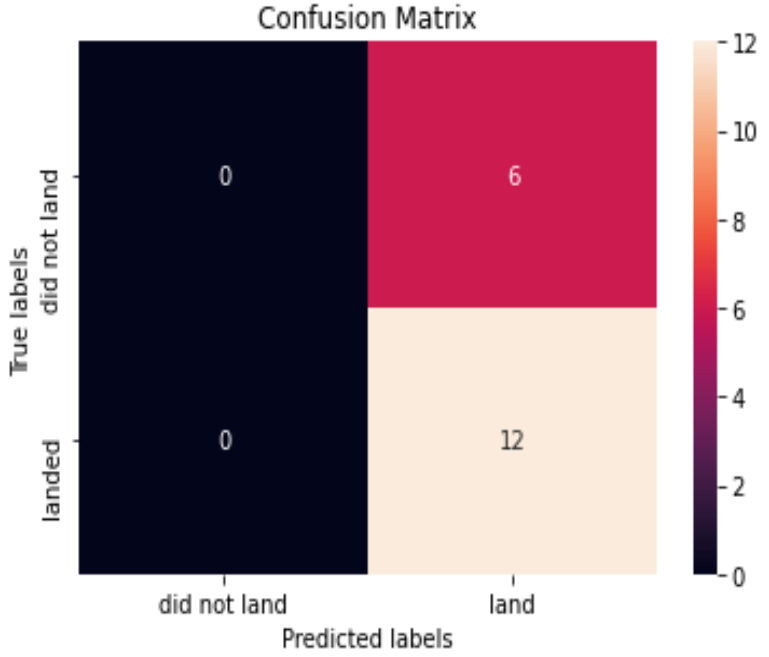
In [31]:
```
algorithms = {'KNN':knn_cv.best_score_, 'SVM':svm_cv.best_score_,'Tree':tree_cv.best_score_,'LogisticRegression':logreg_cv.best_score_}
bestalgorithm = max(algorithms, key=algorithms.get)
print('Best Algorithm is',bestalgorithm,'with a score of',algorithms[bestalgorithm])
if bestalgorithm == 'Tree':
    print('Best Params is :',tree_cv.best_params_)
if bestalgorithm == 'SVM':
    print('Best Params is :',svm_cv.best_params_)
if bestalgorithm == 'KNN':
    print('Best Params is :',knn_cv.best_params_)
if bestalgorithm == 'LogisticRegression':
    print('Best Params is :',logreg_cv.best_params_)
```

```
Best Algorithm is Tree with a score of 0.875
Best Params is : {'criterion': 'gini', 'max_depth': 4, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 5, 'splitter': 'random'}
```

Confusion Matrix

- The confusion matrix is shown on the side. It can be seen that the decision tree can distinguish between different classes. The major problem is that of false positives.

```
In [25]: yhat = tree_cv.predict(X_test)
         plot_confusion_matrix(Y_test,yhat)
```



Conclusions

- The Decision Tree model gave the prediction with the best accuracy
- The Launch Site KSC LC-39A should be used to launch more rockets in the future because it recorded the highest number of success rate amongst the four launch sites.
- The success rate trend for landing outcomes will continue to be positive
- Pay load Mass contribution to success outcomes on the overall orbits is indeterminant; some orbits show positive correlation whereas some do not show any correction at all
- Low weighted payloads perform better than the heavier payloads
- The success rates for SpaceX launches is directly proportional to time. This suggests that as years go by, SpaceX will eventually perfect the launches
- Orbit GEO,HEO,SSO,ES-L1 have the best Success Rate

Appendix

- Find below the link to my github data science repository:
- https://github.com/FestusCJ/Data-Science

Thank you!