IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Debasish Saha
08/06/2022

Outline

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

Executive Summary

- The commercial space age dominated by SpaceX.
- Space X advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars while other providers cost upward of 165 million dollars each, much of the savings is because Space X can reuse the first stage.
- A new rocket launch provider SpaceY wants to bid against SpaceX.
- SpaceY needs to understands why SpaceX is so successful from the data available from SpaceX launches.
- Machine learing algorithms are used to predict the success rate of first stage Falcon 9 rocket.
- SpaceY can use the information to tune its mission parameters and compete against SpaceX to win the space race!

Introduction

Buissiness Problem:
How to predict the outcome of Falcon 9 first stage rocket?

- This report is made to help SpaceY to bid against SpaceX successfully.
- I, as a data scientist on behalf of SpaceY acquire, clean, and prepare the data from SpaceX launches.
- Four supervised machine leaning techniques are used with the data to identify the important mission parameters and predict the success rate.
- This report is a part of Coursera Data Science Professional Certificate (Course # 10: Applied Data Science Capstone).

Section 1

Methodology

Methodology

Executive Summary

- Data collection methodology:
  - Describe how data was collected
- Perform data wrangling
  - Describe how data was processed
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - How to build, tune, evaluate classification models

Data Collection

Data was collected using two methods:

- Data Collection – SpaceX API
- Data Collection - Scraping

Data Collection – SpaceX API

Define helper functions

Request launch data using SpaceX API

Request and parse the data using GET request

Filter the dataframe to only include Falcon 9 launches

Source code: https://github.com/debusaha/DS_Capstone/blob/main/SpaceX_Data_Collection_API.ipynb

Data Collection - Scraping

Define some helper functions to process web scraped HTML table

Request the Falcon 9 wiki HTML page from its URL and get a HTML response object

Create a BeautifulSoup object from the HTML response object

Extract all column/variable names from the HTML table header

Create a data frame by parsing the launch HTML tables

Source code: https://github.com/debusaha/DS_Capstone/blob/main/SpaceX_Data_Collection_with_web_scraping.ipynb

Data Wrangling

Perform exploratory data analysis

↓

Calculate the number of launches on each site

↓

Calculate the number and occurrence of each orbit

↓

Calculate the number and occurrence of mission outcome per orbit type

↓

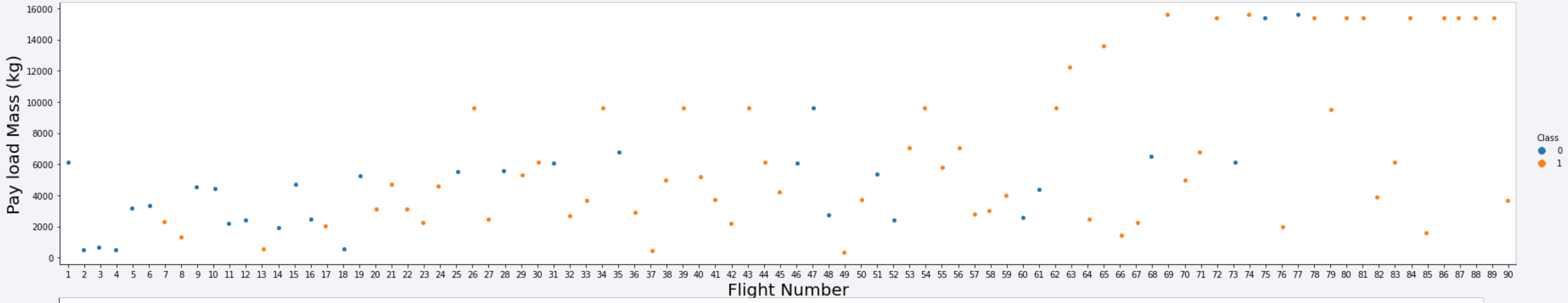
Determine training labels: Create a landing outcome label from outcome column

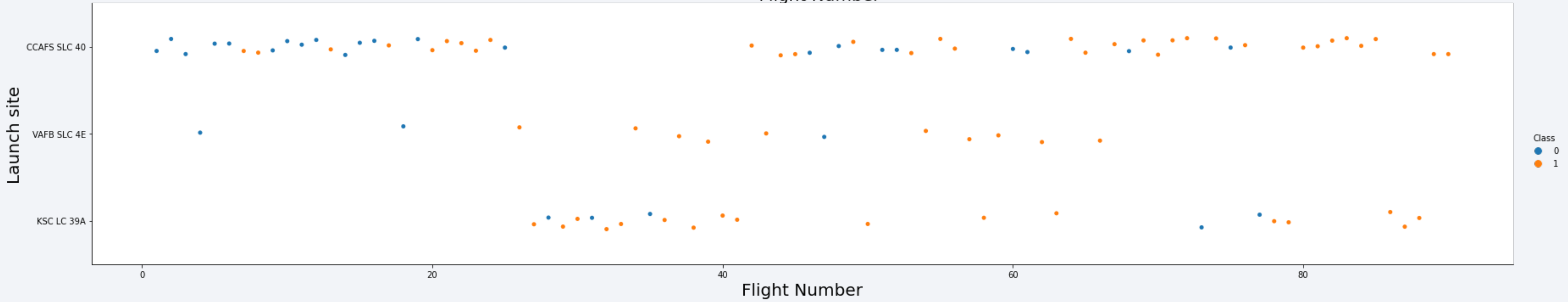
Source code: https://github.com/debusaha/DS_Capstone/blob/main/SpaceX_Data_Wrangling.ipynb

EDA with Data Visualization

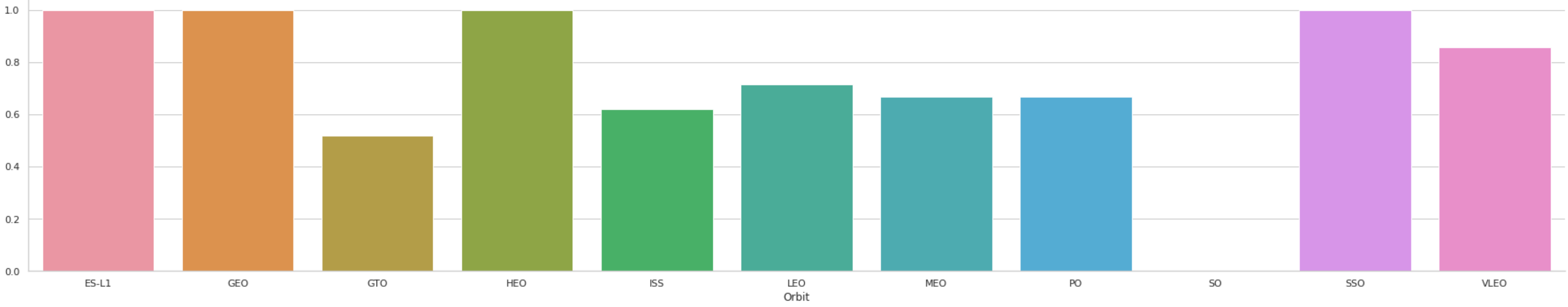
Following graphs are plotted using Matplotlib and Seaborn:

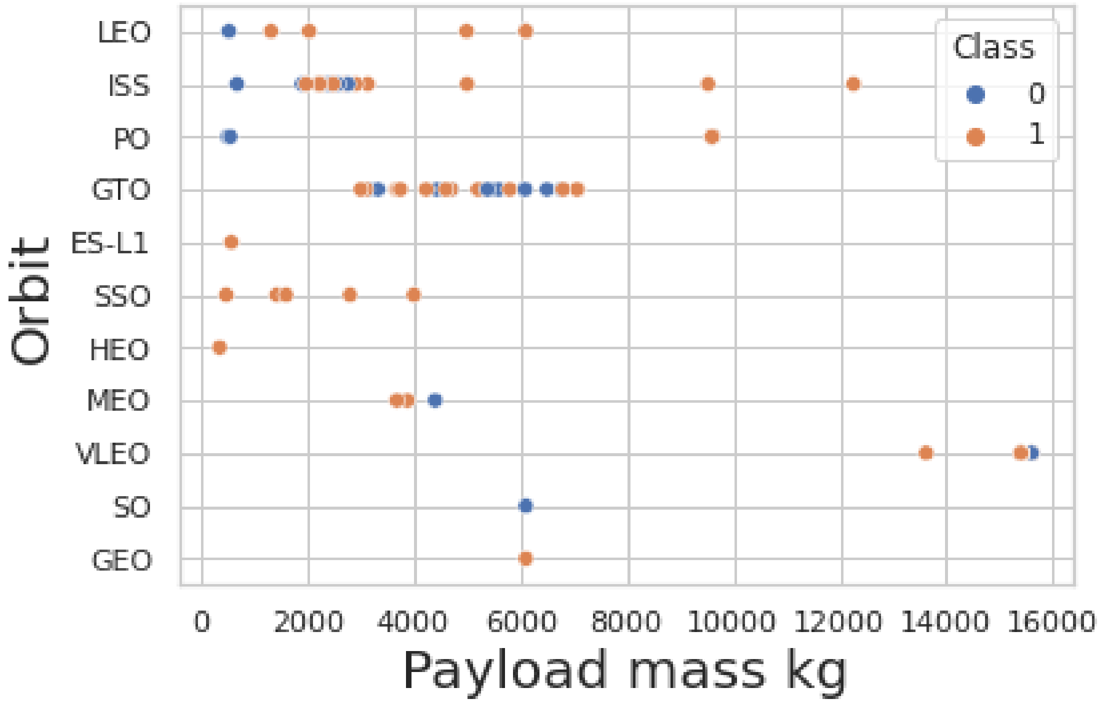
1. Payload mass vs. Flight Number
2. Launch Site vs. Flight Number
3. Success rate vs. Orbit type (bar plot)
4. Orbit type vs. Payload mass
5. Orbit type vs. Flight Number

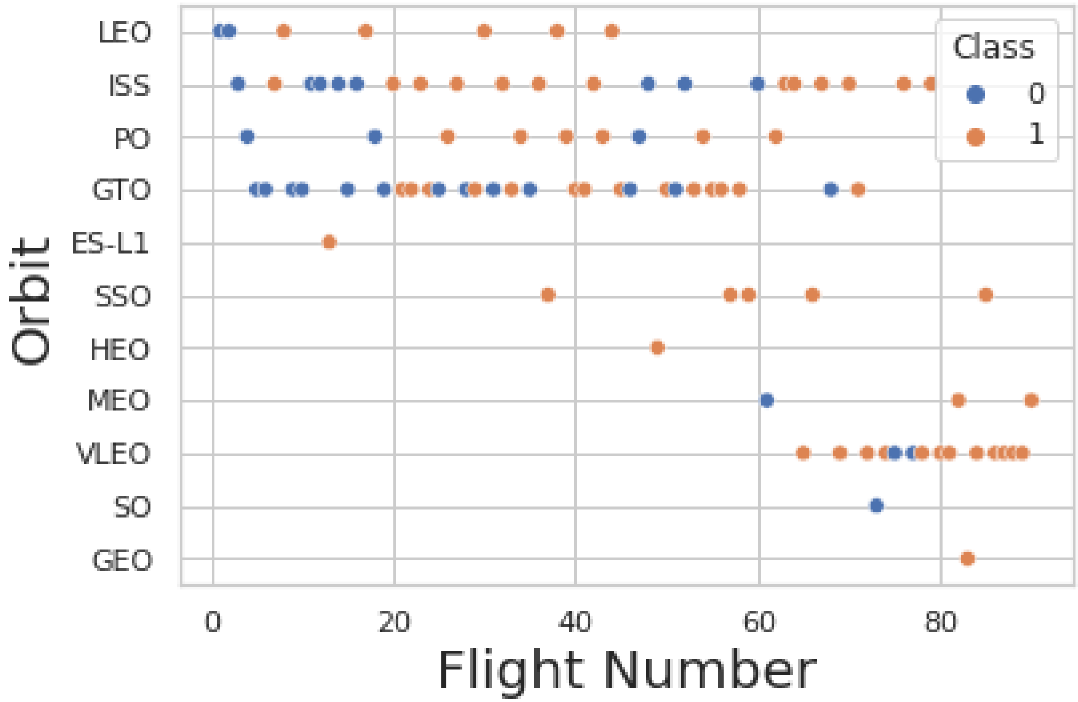




EDA with Data Visualization







Source code: https://github.com/debusaha/DS_Capstone/blob/main/EDA_with_Visualization.ipynb

EDA with SQL

Following SQL queries are performed:

- Display the names of the unique launch sites in the space mission
- Display 5 records where launch sites begin with the string ‘CCA’
- Display the total payload mass carried by boosters launched by NASA (CRS)
- Display average payload mass carried by booster version F9 v1.1
- List the date when the first successful landing outcome in ground pad was achieved
- List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
- List the total number of successful and failure mission outcomes
- List the names of the booster version which have carried the maximum payload mass
- List the failed landing_outcome in drone ship, their booster versions, and launch site names for in year 2015
- Rank the count of landing outcomes between the date 2010-06-04 and 2017-03-20, in descending order

Source code: https://github.com/debusaha/DS_Capstone/blob/main/EDA_with_SQL.ipynb

Build an Interactive Map with Folium

Following map objects are created and added to a folium map.

- Markers: to indicate points e.g. launch sites
- Circles: to indicate highlighted area surrounding a specific coordinate
- Marker clusters: to indicate groups of events e.g. launches in a launch site
- Lines: to measure distances between two coordinates

Source code:
https://github.com/debusaha/DS_Capstone/blob/main/SpaceX_Interactive_Visual_Analytics_with_Folium%20.ipynb

Build a Dashboard with Plotly Dash

Following plots/graphs and interactions are added to a dashboard:

1. Pie chart showing percentage of launches by site: This helps us to analyze the role of launch site for a successful launch.

2. Range of payload mass for each site: This shows the role of payload mass for successful launch.

Source code: https://github.com/debusaha/DS_Capstone/blob/main/SpaceX_dash.py

Predictive Analysis (Classification)

1. Data preparation
2. Data standardization
3. Modeling train data with classification models (kNN, Logistic Regression, Decision Tree and Support Vector Machine)
4. Model the test data using trained model
5. Compare the accuracy of each model and choose the model with best accuracy

Source code: https://github.com/debusaha/DS_Capstone/blob/main/SpaceX_Machine_Learning_Prediction.ipynb

Results

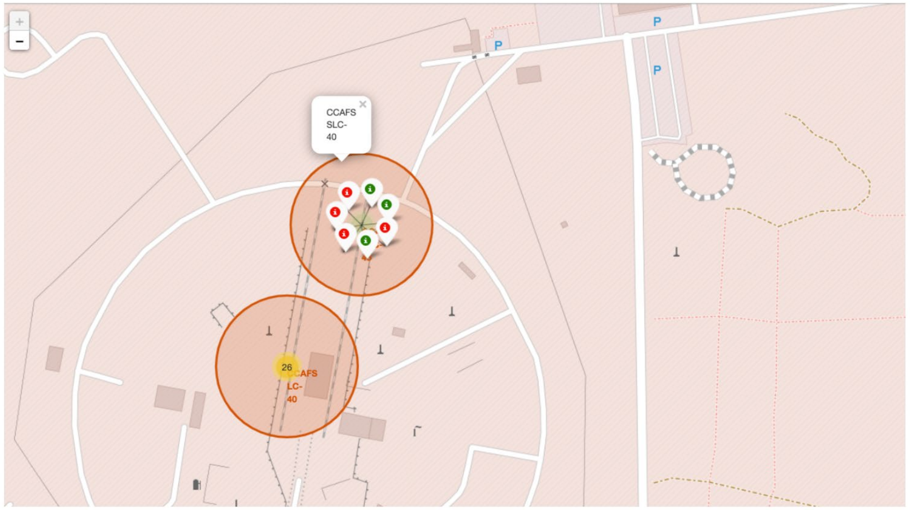
Exploratory data analysis results:

- As the flight number increases, the first stage is more likely to land successfully.
- More massive the payload, the less likely the first stage will return.
- Different launch sites have different success rates.
- No rocket launched from VAFB-SLC launchsite with heavy payload (>10000 kg).
- ES-L1, GEO, HEO and SSO orbits have very high success rate.
- Polar, LEO and ISS have more successful landing with heavy payloads.
- Success rate is increasing since 2013.

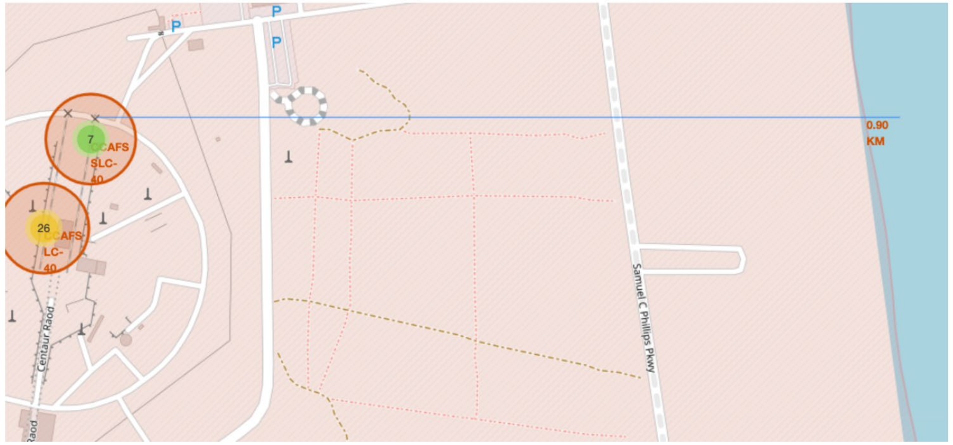
Results

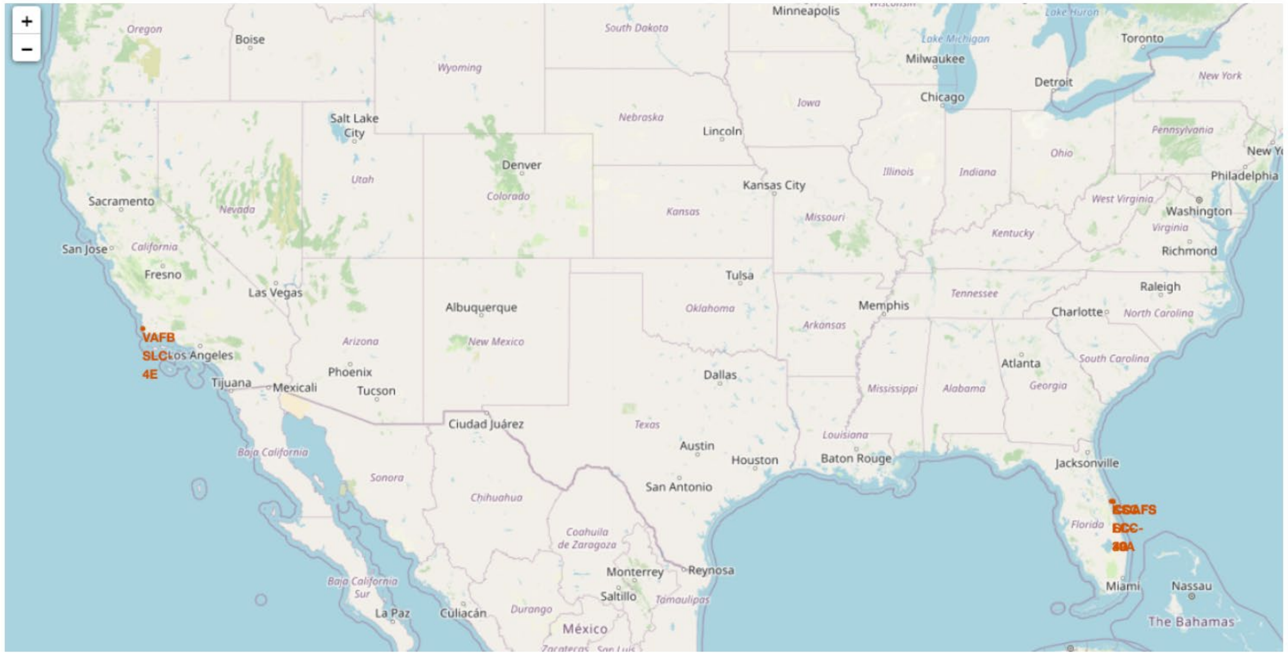
- Interactive analytics demo in screenshots

Mark the success/failed launches for each site on the map



Calculate the distances between a launch site to its proximities





Mark all launch sites on a map

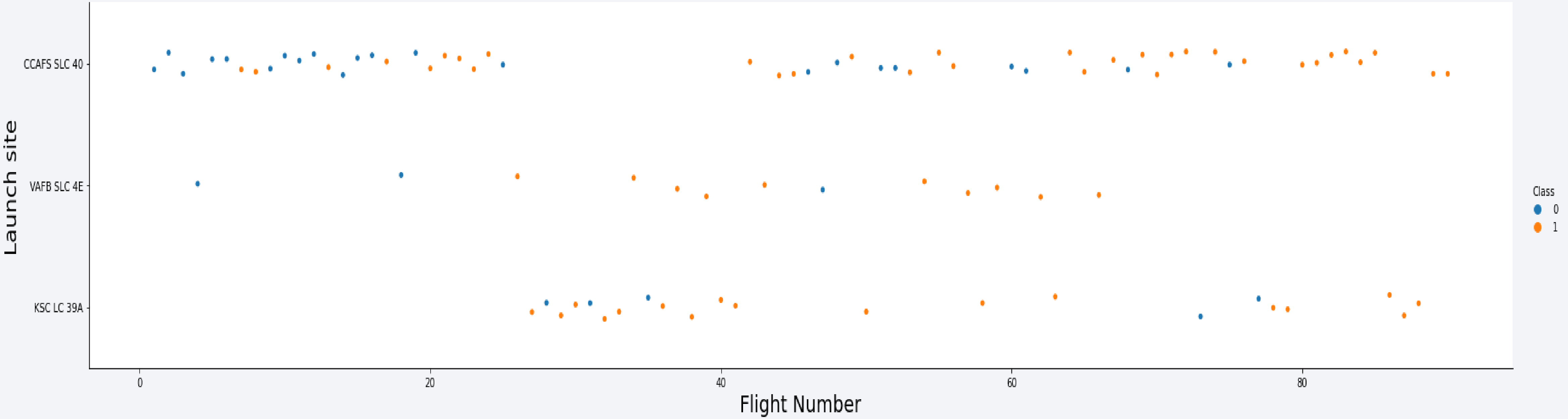
Predictive analysis results

Classification Algorithm	Accuracy (Train data)	Accuracy (Test data)
Logistic Regression	84.6%	83.3%
Support Vector Machine	84.8%	83.3%
Decision Tree	**87.5%**	83.3%
k Nearest Neighbors	84.8%	83.3%

Section 2

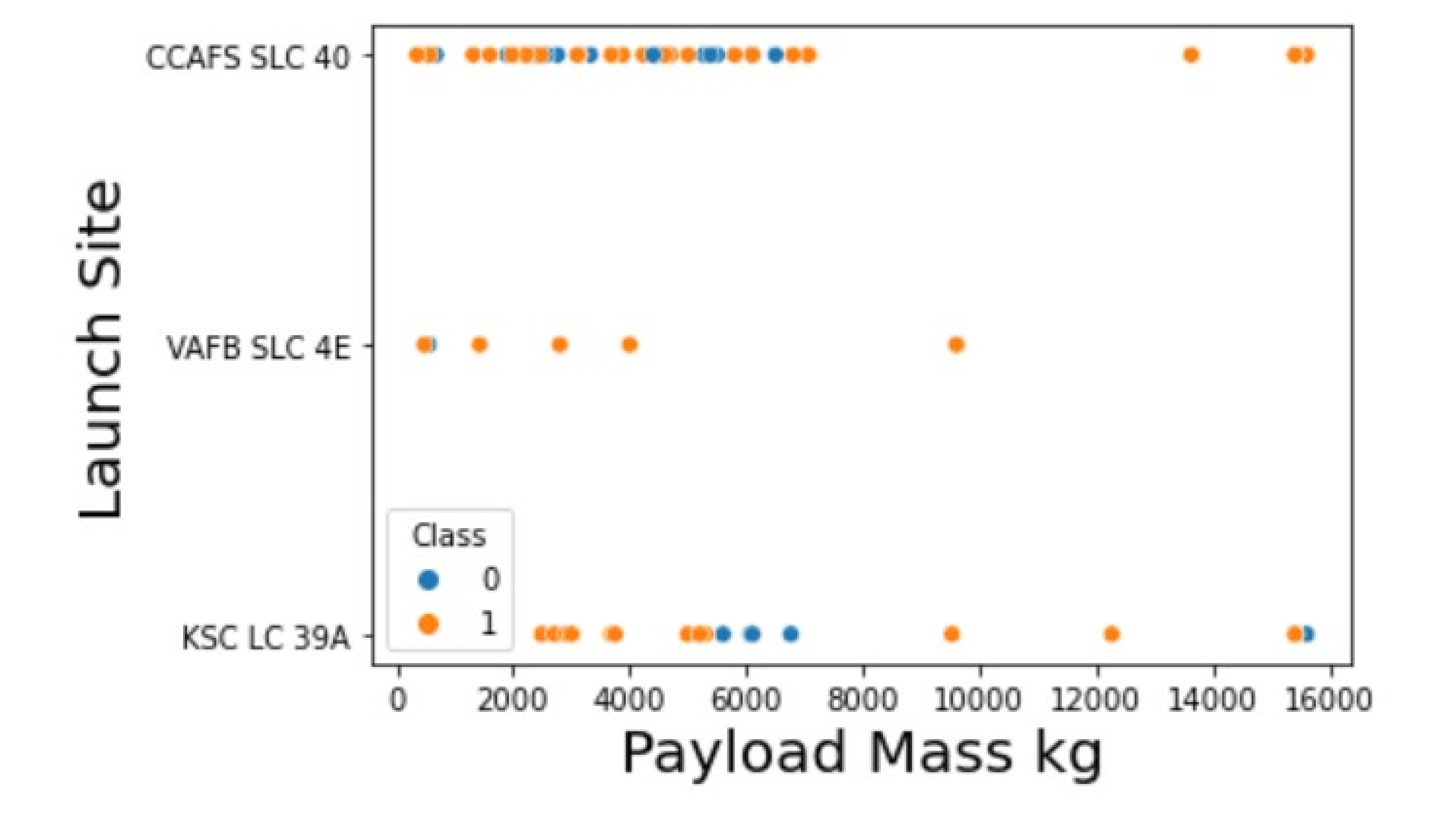
Insights drawn from EDA

Flight Number vs. Launch Site



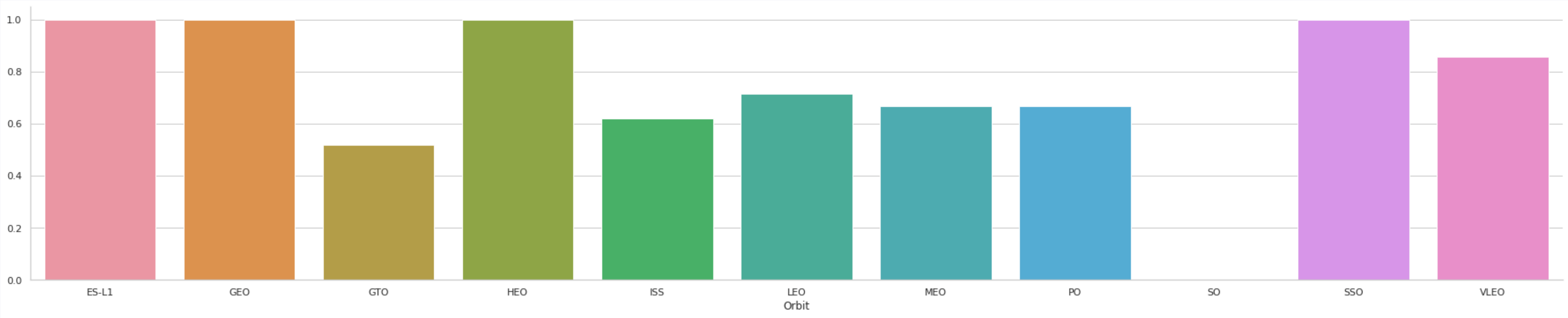
- Success rate improved over time.
- CCAF5 SLC 40 has best success rate.

Payload vs. Launch Site



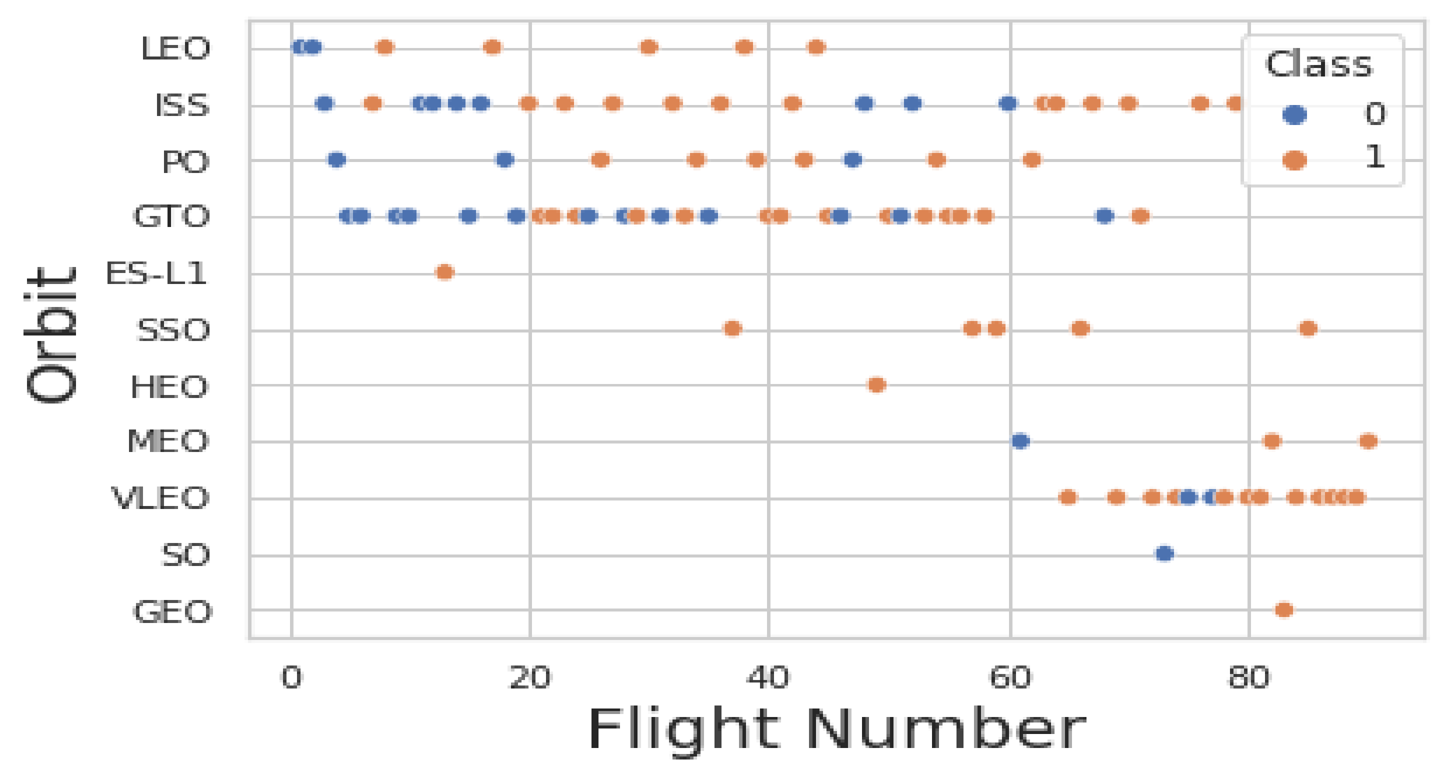
- Very high success rate for payload mass > 8000 kg

Success Rate vs. Orbit Type



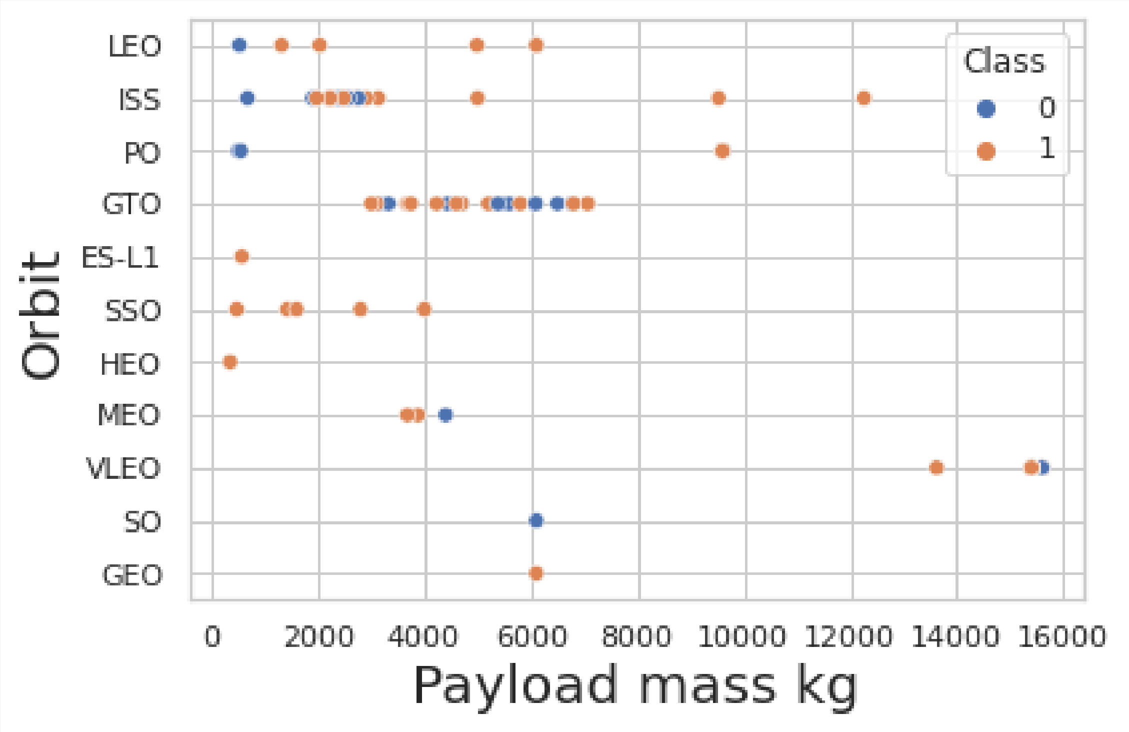
- These four orbit types (ES-L1, GEO, HEO and SSO) have very high success rate.

Flight Number vs. Orbit Type



- Recent launches are more successful and VLEO and ISS orbit types have high recent success rate.

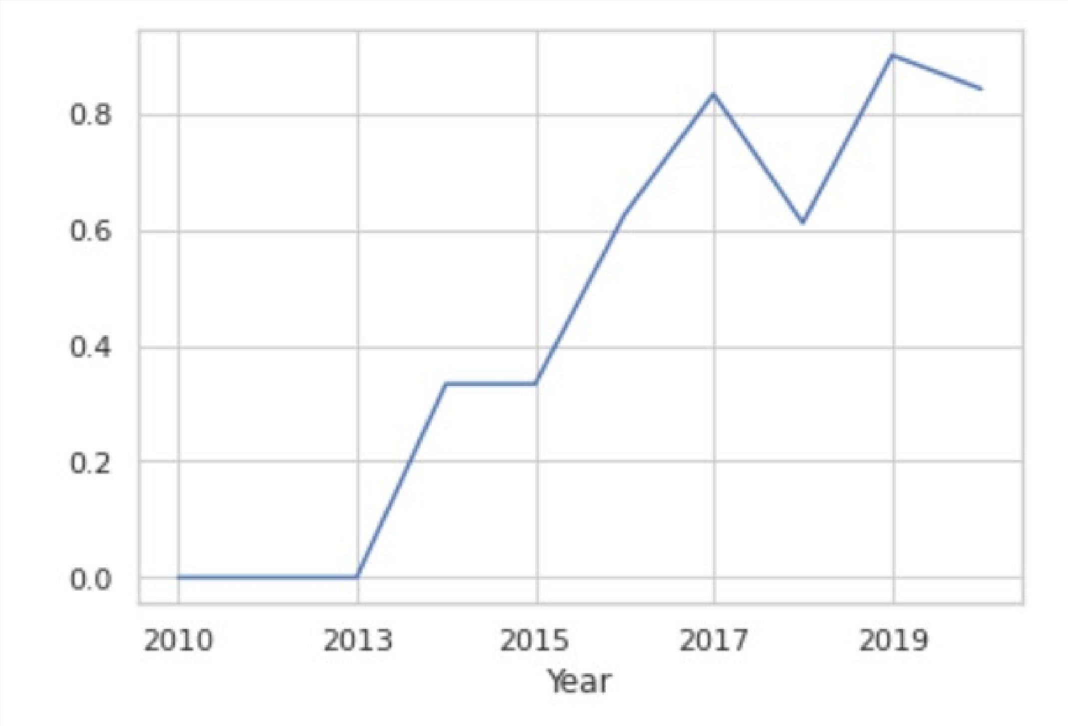
Payload vs. Orbit Type



- GTO and ISS have wide variety of payloads.
- SSO has 100% success rate for all payloads.
- SO, GEO, ES-L1 have only one payload with 100% success rate.

Launch Success Yearly Trend

- yearly average success rate improves with time.



;

All Launch Site Names

Names of the unique launch sites:

1. CCAFS SLC 40
2. CCAFS LC 40
3. KSC LC 39A
4. VAFB SLC 4E

This is obtained by the following SQL command:

sql SELECT DISTINCT LAUNCH_SITE FROM SPACEXTBL ORDER BY 1;

Launch Site Names Begin with 'CCA'

Find 5 records where launch sites begin with `CCA`:

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

This is obtained by the following SQL command:

sql SELECT * FROM SPACEXTBL WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5;

Total Payload Mass

Calculate the total payload carried by boosters from NASA:

total_payload
111268

This is obtained by the following SQL command:

sql SELECT SUM(PAYLOAD_MASS__KG_) AS TOTAL_PAYLOAD FROM SPACEXTBL WHERE PAYLOAD LIKE '%CRS%';

Average Payload Mass by F9 v1.1

Calculate the average payload mass carried by booster version F9 v1.1

avg_payload
2928

This is obtained by the following SQL command:

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

First Successful Ground Landing Date

Find the dates of the first successful landing outcome on ground pad

first_success_gp
2015-12-22

This is obtained by the following SQL command:

sql SELECT MIN(DATE) AS FIRST_SUCCESS_GP FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Success (ground pad)';

Successful Drone Ship Landing with Payload between 4000 and 6000

List the names of boosters which have successfully landed on drone ship and had payload mass greater than 4000 but less than 6000

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

This is obtained by the following SQL command:

sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000 AND LANDING__OUTCOME = 'Success (drone ship)';

Total Number of Successful and Failure Mission Outcomes

Calculate the total number of successful and failure mission outcomes

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

This is obtained by the following SQL command:

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

Boosters Carried Maximum Payload

List the names of the booster which have carried the maximum payload mass

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

This is obtained by the following SQL command:

sql SELECT DISTINCT BOOSTER_VERSION FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL) ORDER BY BOOSTER_VERSION;

2015 Launch Records

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

booster_version	launch_site
F9 v1.1 B1012	CCAFS LC-40
F9 v1.1 B1015	CCAFS LC-40

This is obtained by the following SQL command:

sql SELECT BOOSTER_VERSION, LAUNCH_SITE FROM SPACEXTBL WHERE LANDING__OUTCOME = 'Failure (drone ship)' AND DATE_PART('YEAR', DATE) = 2015;

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

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

landing__outcome	qty
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

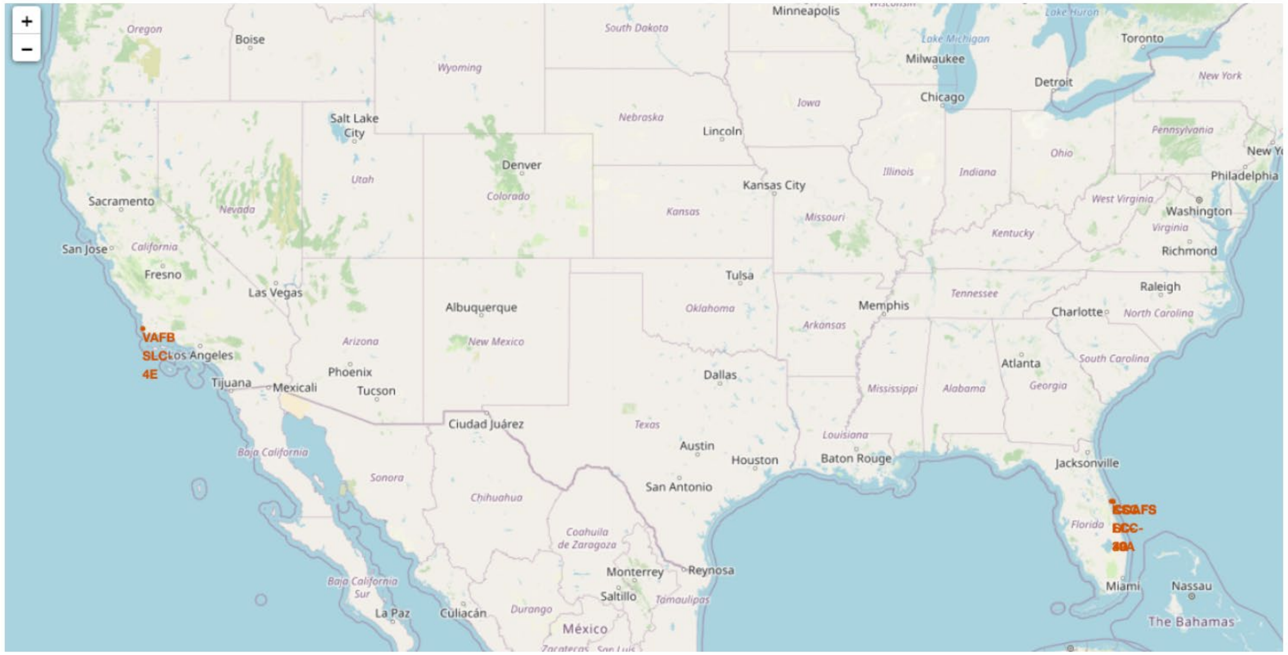
This is obtained by the following SQL command:

sql SELECT LANDING__OUTCOME, COUNT(*) AS QTY FROM SPACEXTBL WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20' GROUP BY LANDING__OUTCOME ORDER BY QTY DESC;

Section 3

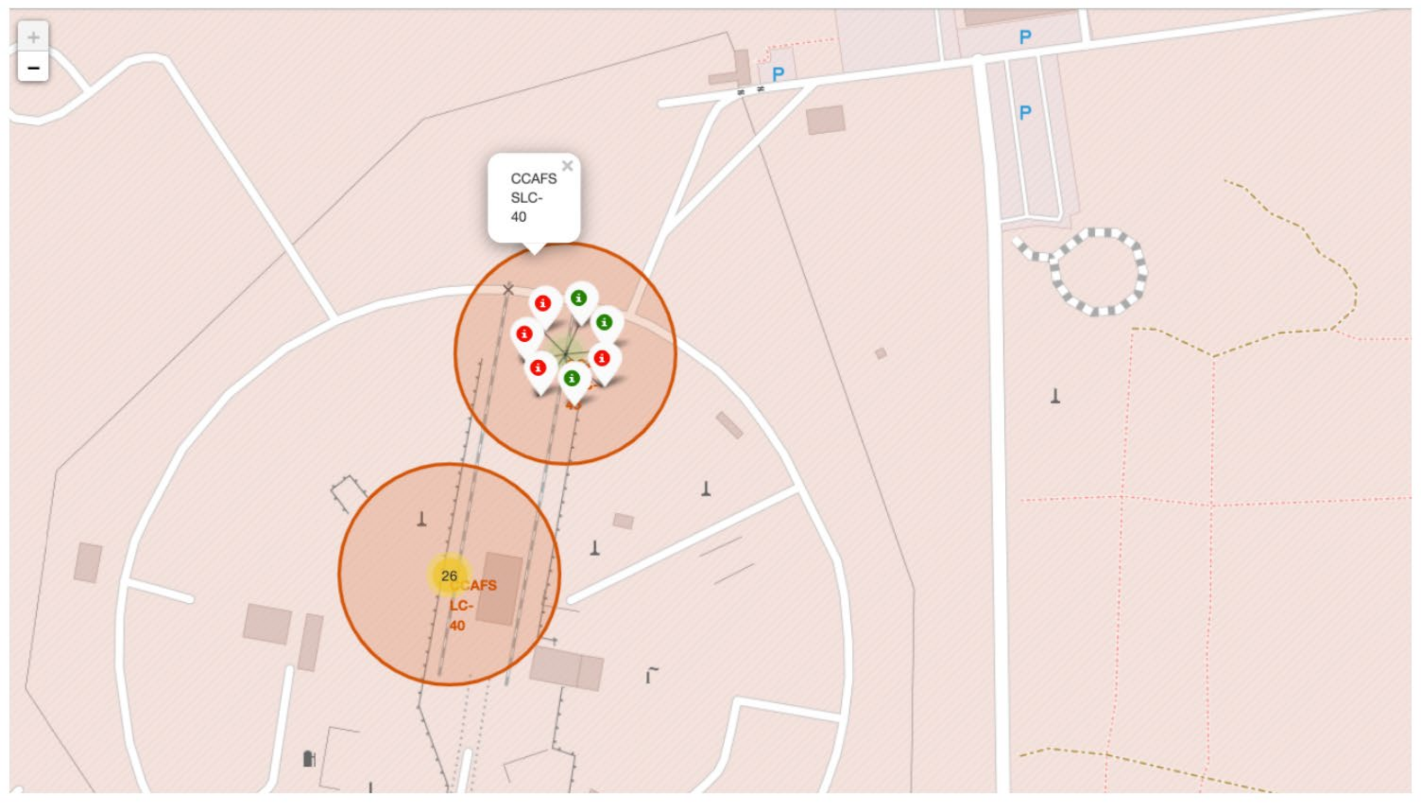
Launch Sites Proximities Analysis

All Launch sites



All launch sites are in very close proximity to the coast.

Outcome of Launch

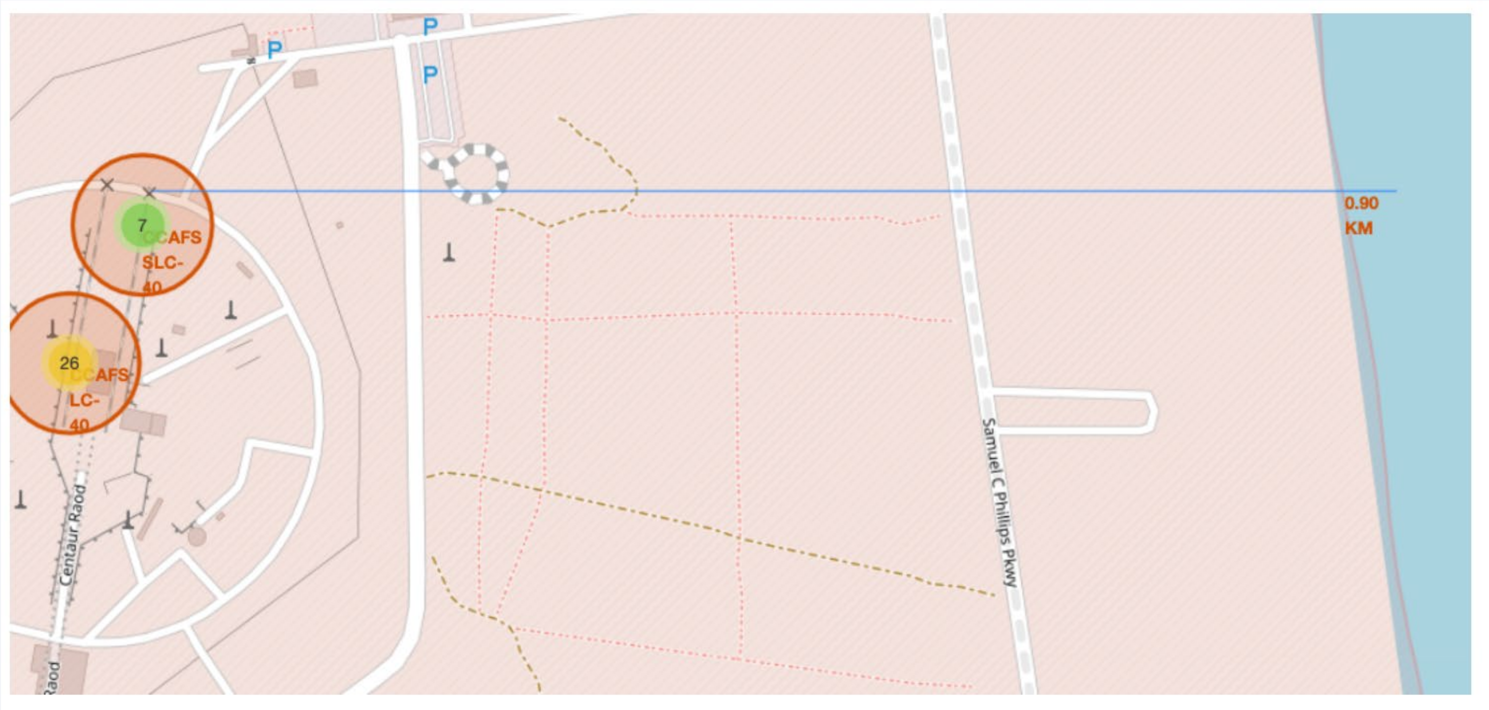


Launch site outcome of CCAFS SLC-40

- Green marker: Success
- Red marker: Failure

Launch site to its proximities

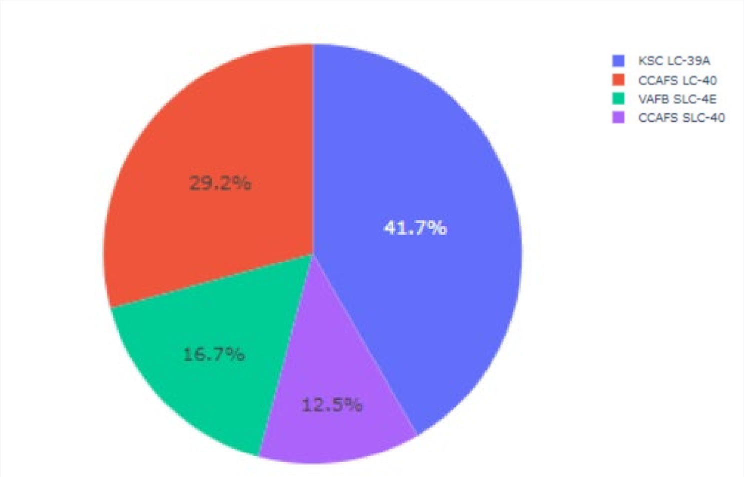
Blue line shows the distance between launch site CCAFS SLC-40 and coastline.



Section 4

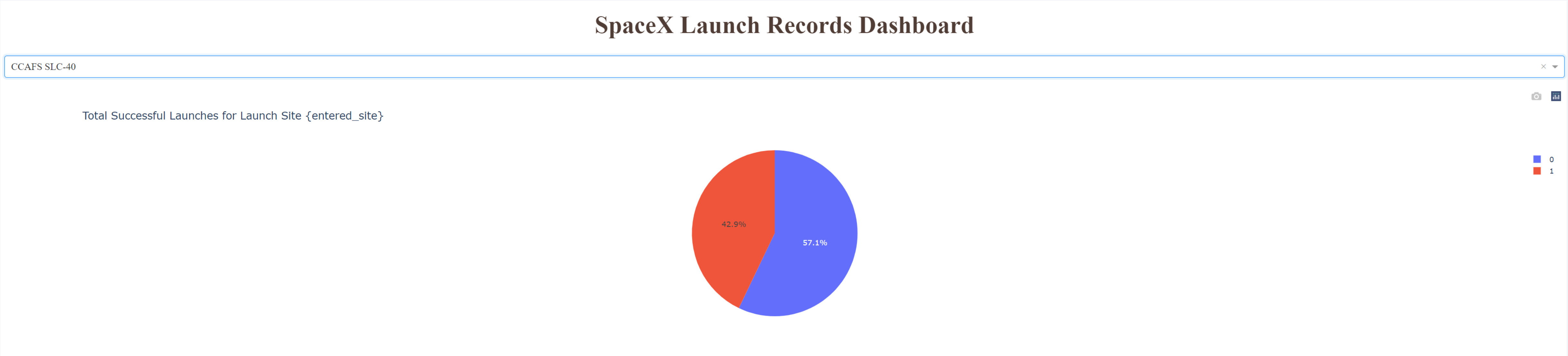
Build a Dashboard with Plotly Dash

Launch success count for all sites



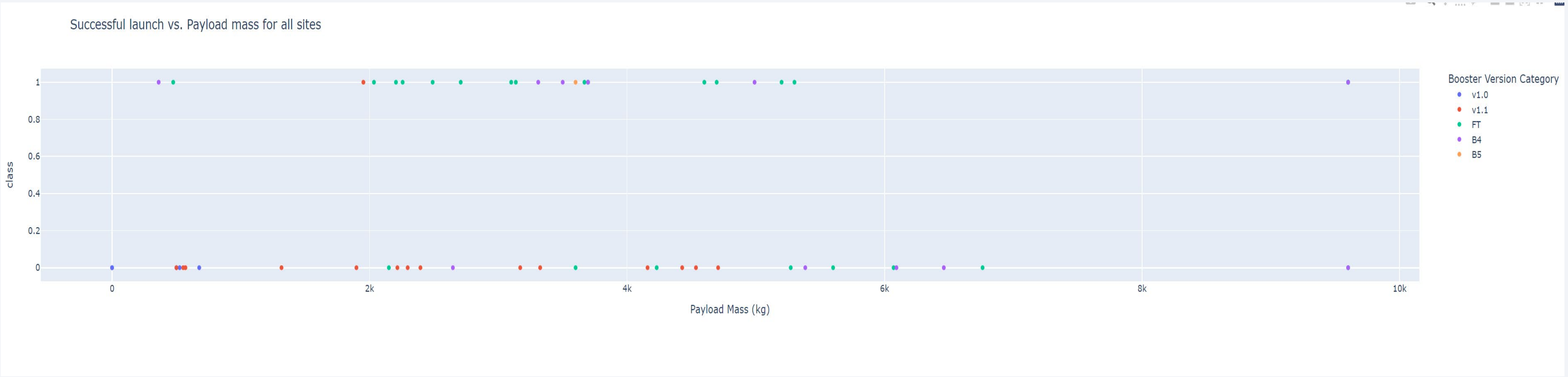
CCAFS SLC-40 is the most successful launch site.

Most successful launch site



CCAFS SLC-40 launch site has highest launch success ratio (42.9%).

Payload vs. Launch Outcome

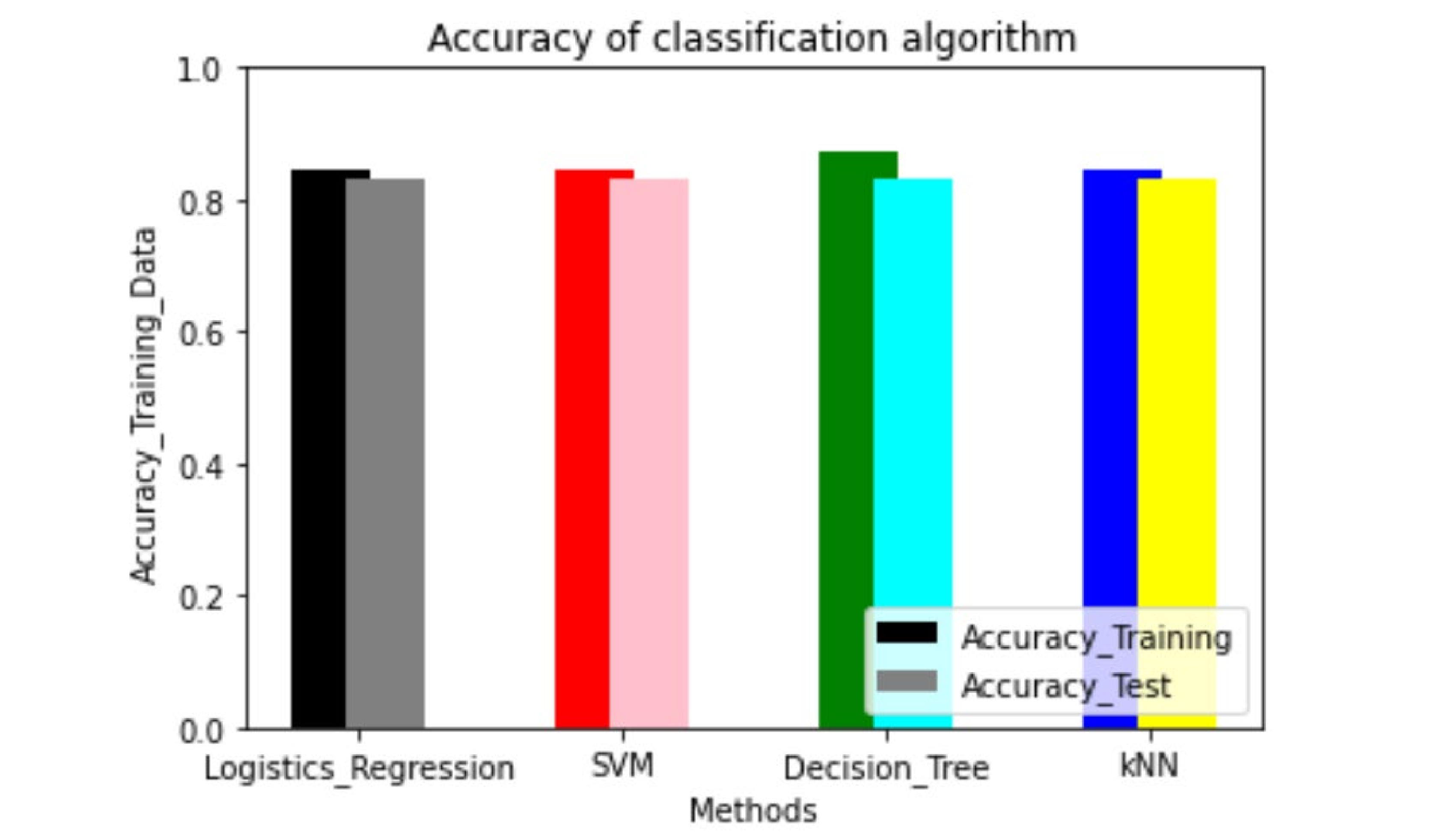


Successful launch outcome depends on payload. Launches are more successful for payload mass < 6000 kg.

Section 5

Predictive Analysis (Classification)

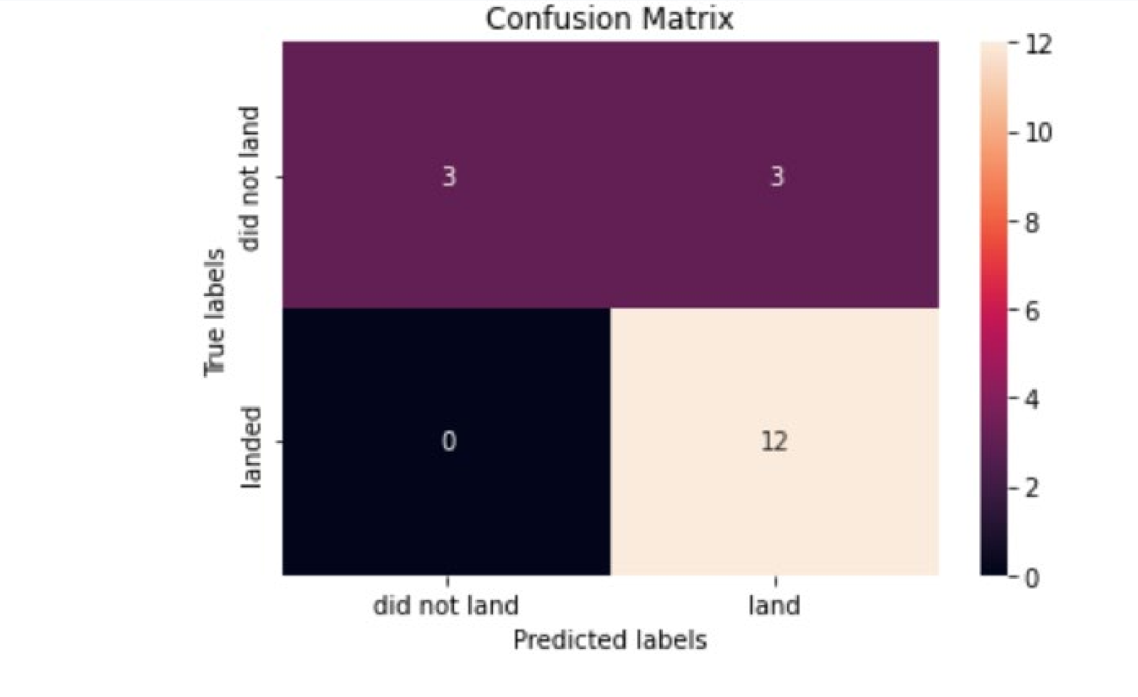
Classification Accuracy



- Decision Tree has the highest classification accuracy

Confusion Matrix

Decision Tree is the most successful model.



Decision Tree detected fewer false negative cases (only 3). Therefore, they have highest accuracy.

Conclusions

1. CCAFS SLC-40 is more suitable for launching rocket.
2. Launch success improves over time.
3. Decision Tree model is best to predict the launch outcome.

Appendix

- https://github.com/debusaha/DS_Capstone/
- https://en.wikipedia.org/wiki/SpaceX

Thank you!