IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

Mariana Bujac-Leisz

July 18th, 2023

Outline

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

Executive Summary

- Summary of methodologies
  - Data Collection - Webscraping and SpaceX API
  - Data Wrangling - replace missing values
  - Exploratory Data Analysis -
    - Analyze outcomes with SQL
    - Visual Analysis
  - Interactive Dashboard
  - Prediction Analysis using Classification
    - Logistic Regression, SVM, Decision Tree, KNN
- Summary of all results
  - Launch success rate increases over time
  - Higher success rate for higher orbits
  - Higher success rate for higher payload mass
  - Lower success rate for booster versions v1.0, v1.1; higher success rate for FT, B4, B5
  - Higher success rate for Kennedy Space center and at Cape Canaveral

Introduction

- Project background and context
  - SpaceX advertises low-cost Falcon 9 rocket launches (average of $62m vs. $165m of competitors).
  - This success is because of the reusability of the first stage.
- Problems you want to solve
  - If we can determine that the first stage will land, we can determine the cost of a launch.
  - This information can be used if an alternative company wants to bid against SpaceX for a rocket launch.

https://github.com/mbujac/Applied-Data-Science-Capstone/blob/ca579780d9b7a7e2ffd9f0f616e7bda7cf9c228a/README.md

Section 1

Methodology

Methodology

Executive Summary

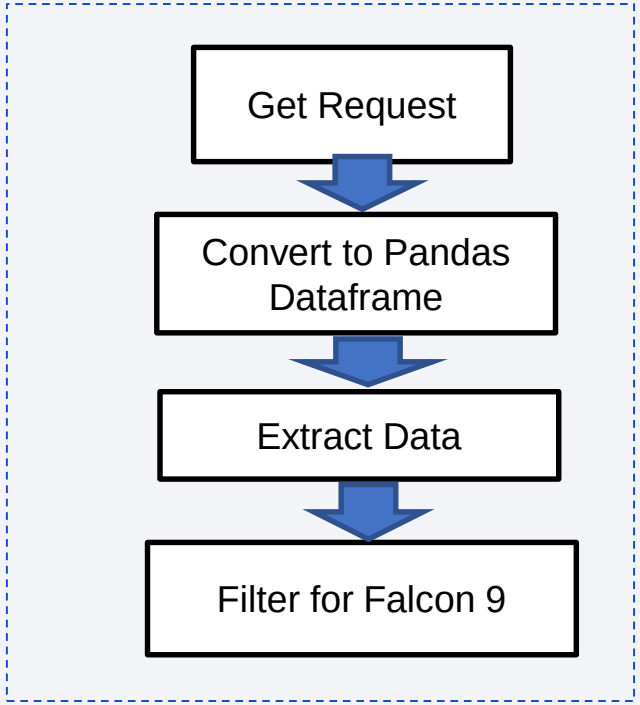
- Data collection methodology: SpaceX-API; Webscraping of SpaceX Wikipedia page
- Perform data wrangling – Convert some outcomes into Training Labels
- Perform exploratory data analysis (EDA) using visualization and SQL
    - Analyze outcomes by orbit type, payload mass, and booster versions with SQL
    - Visual Analysis with charts by payload mass, time, orbit type and launch site
- Perform interactive visual analytics using Folium and Plotly Dash
    - Visual Analysis with maps by site
    - Interactive Dashboard - analysis by site, payload and booster
- Perform predictive analysis using classification models
    - Logistic Regression, SVM, Decision Tree, KNN
    - Parameter tuning with Grid Search

Data Collection

- SpaceX REST API
  - RESTful Interface
  - Get Core Data, Booster Version, Launch Site Data, Payload Data
- Webscaping of SpaceX Wikipedia Page
  - HTML Requests
  - Python’s BeautifulSoup package for webscraping
  - Extract Column names from HTML table header

Data Collection – SpaceX API

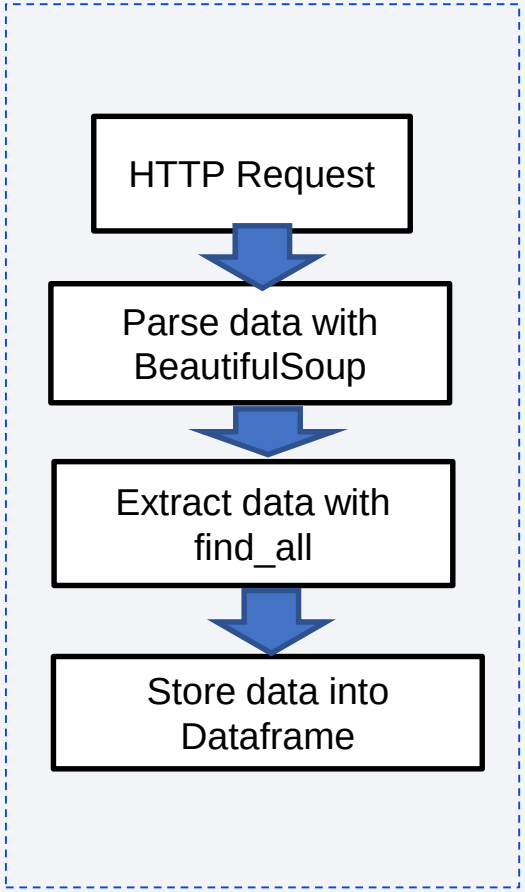
- Send Get Request to SpaceX API interface website
- Parse data into a Pandas dataframe
- Extract data with specific functions for:
  - Core data
  - Launch Site Data
  - Payload Mass
  - Booster Version
- Filter data for Falcon 9
- Data Collection (RESTful API) - https://github.com/mbujac/Applied-Data-Science-Capstone/blob/ca579780d9b7a7e2ffd9f0f616e7bda7cf9c228a/1-data-collection-api.ipynb



Data Collection - Scraping

- Send HTTP Request to SpaceX Wikipedia website
- Parse data into Pandas dataframe with BeautifulSoup
- Extract data with find_all method
- Store data into Pandas dataframe for further use

Data Collection (Webscraping) - https://github.com/mbujac/Applied-Data-Science-Capstone/blob/ca579780d9b7a7e2ffd9f0f616e7bda7cf9c228a/2-webscraping.ipynb



Data Wrangling

- Create Training Labels for Outcome column with the value of 1 when the booster successfully landed and 0 if it was unsuccessful.

- Data Wrangling - https://github.com/mbujac/Applied-Data-Science-Capstone/blob/ca579780d9b7a7e2ffd9f0f616e7bda7cf9c228a/3-spacex-data_wrangling.ipynb

EDA with SQL

- Display the names of 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 recors which will display the month names, failure landing_outcomes in drone ship, booster_versions, launch_site for the months in year
- 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

Data Analysis with SQL - https://github.com/mbujac/Applied-Data-Science-Capstone/blob/ca579780d9b7a7e2ffd9f0f616e7bda7cf9c228a/4-eda-sqllite.ipynb

EDA with Data Visualization

We used the following charts:

- Flight number vs Payload mass – as the flight number increases, the first stage is more lakely to land successfully
- Relationship between Flight number and Launch site – shows the success rate of each launch site over time
- Payload mass vs Launch site – shows which payload is successful at each launch site
- Orbit type vs. Success rate – shows which orbit types have the highest success rates
- Orbit type vs. Flight number – shows the development of orbit types over time
- Orbit type vs. Payload mass – shows the success rate for specific orbit type for a payload mass
- Success rate vs. Year – shows rate since 2013 kept increasing over time

Data Exploration – https://github.com/mbujac/Applied-Data-Science-Capstone/blob/ca579780d9b7a7e2ffd9f0f616e7bda7cf9c228a/5-eda-dataviz.ipynb

Build an Interactive Map with Folium

- Mark all launch sites on a map
- Mark the success/failed launches for each site on the map
- Calculate the distances between a launch site to its proximities

Interactive Map with Folium – https://github.com/mbujac/Applied-Data-Science-Capstone/blob/ca579780d9b7a7e2ffd9f0f616e7bda7cf9c228a/6-launch_site_location-Interactive-Visual-Analytics-with-Folium.ipynb

Build a Dashboard with Plotly Dash

Input Elements:

- Dropdown list for the launch site (with select all as one of the options)
- RangeSlider for selecting the payload mass

Output Elements:

- PieChart: for showing the success rate of each launch site, or the number of successful landing outcomes for a landing site
- Scatterplot: Show success/failure by payload and booster version

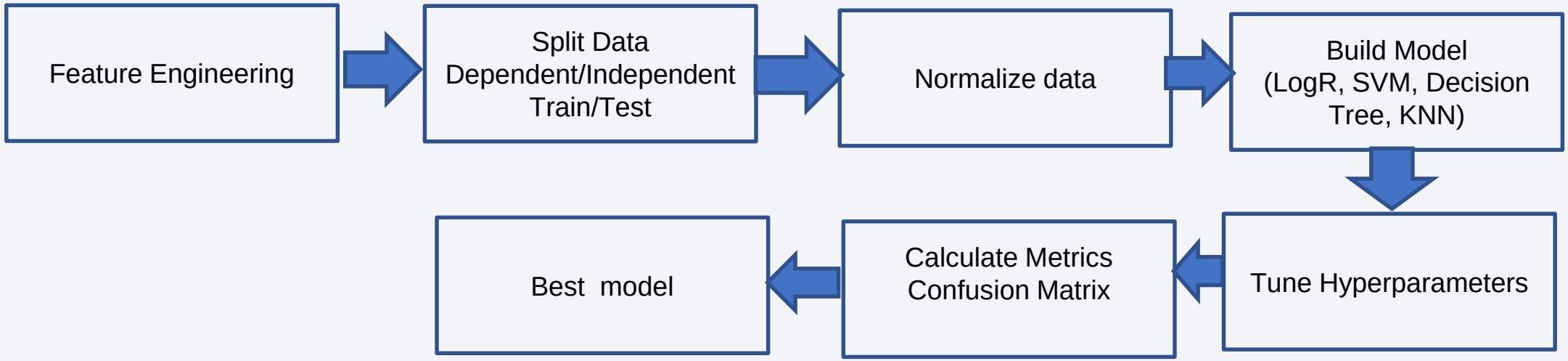
Interactive Dashboard with Plotly – https://github.com/mbujac/Applied-Data-Science-Capstone/blob/ca579780d9b7a7e2ffd9f0f616e7bda7cf9c228a/7-build_a_dashboard_application_with_Plotly_Dash.ipynb

Predictive Analysis

- Preprocessing
  - One-Hot-Encoding for Categorical Features
  - Split data into dependent/independent variables and train/test data
  - Scale Data with StandardScaler
- Model Building for each Method
  - Logistic Regression
  - Support Vector Machine
  - Decision Tree
  - K-Nearest Neighbor
- Optimization
  - Use Gridsearch for tuning the hyperparameters
  - Examining the Confusion Matrix
- Evaluation
  - Use Accuracy of Gridsearch for selecting the best parameter
  - Use Score to compare each classification method

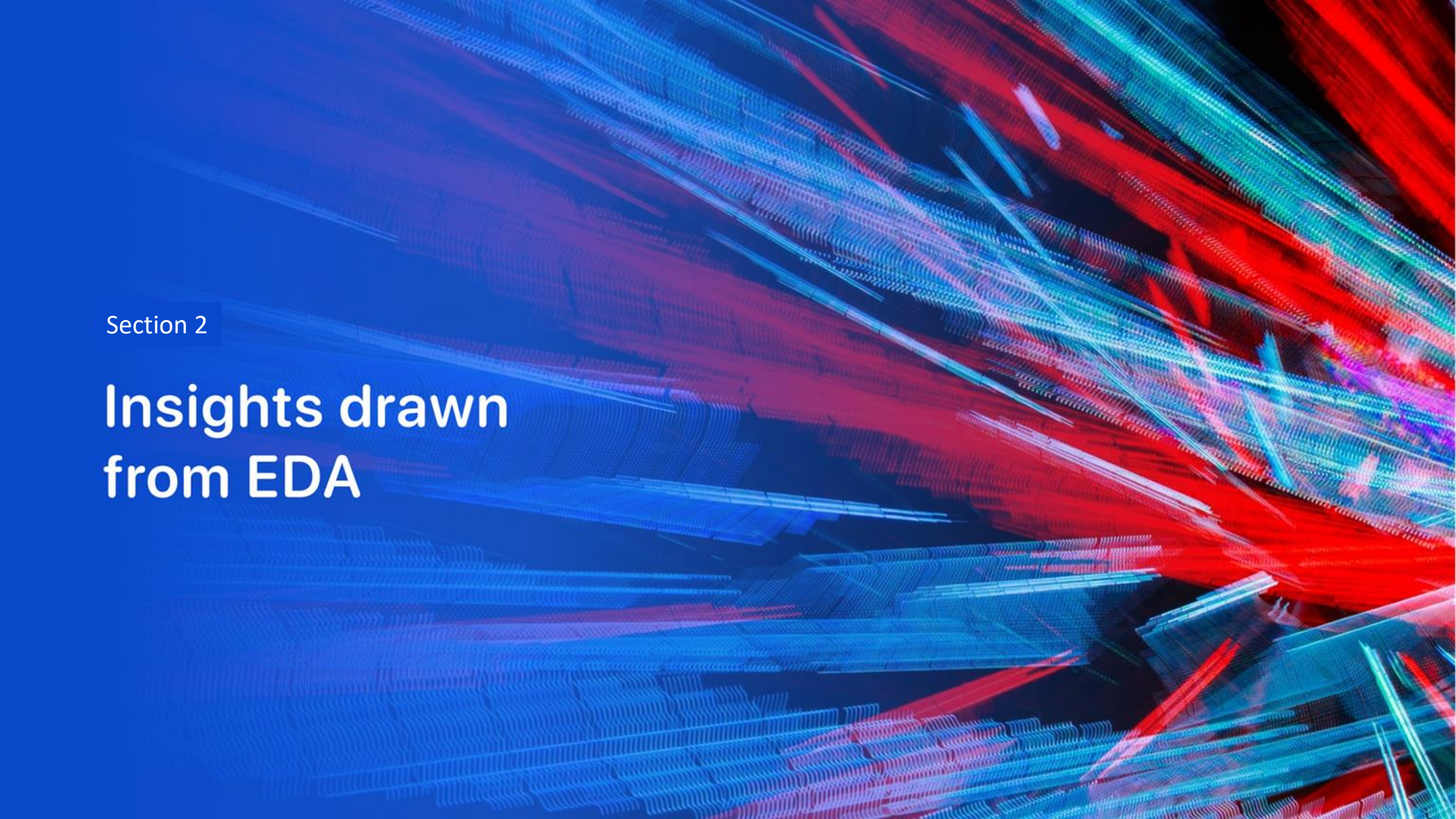
- Machine Learning Prediction – https://github.com/mbujac/Applied-Data-Science-Capstone/blob/ca579780d9b7a7e2ffd9f0f616e7bda7cf9c228a/8-SpaceX_Machine_Learning_Prediction.ipynb

Predictive Analysis

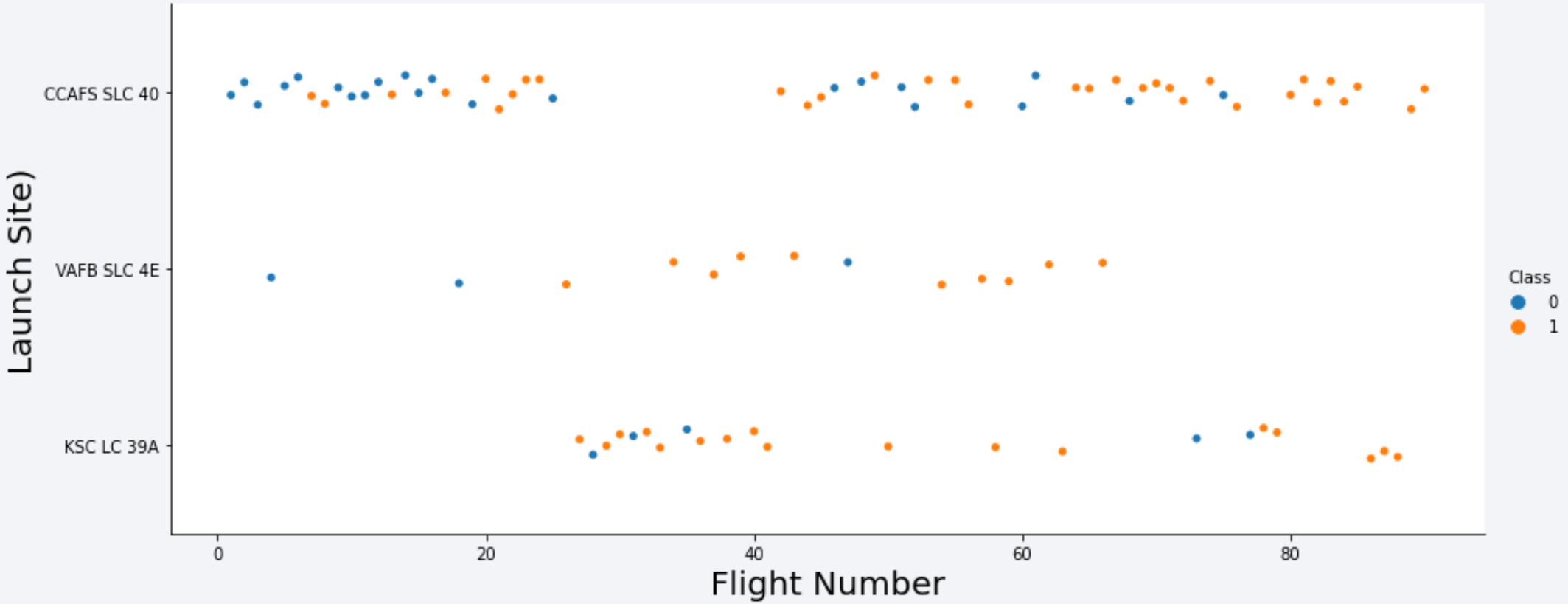


Results

- Launch success rate increases over time
- Higher success rate for higher orbits
- Higher success rate for higher payload mass
- Low success rate for booster versions v1.0, v1.1, high success rate for FT, B4, B5
- Higher success rate for Kennedy Space center and recent starts at Cape Canaveral
- Best prediction results with Logistic Regression and Support Vector Machine

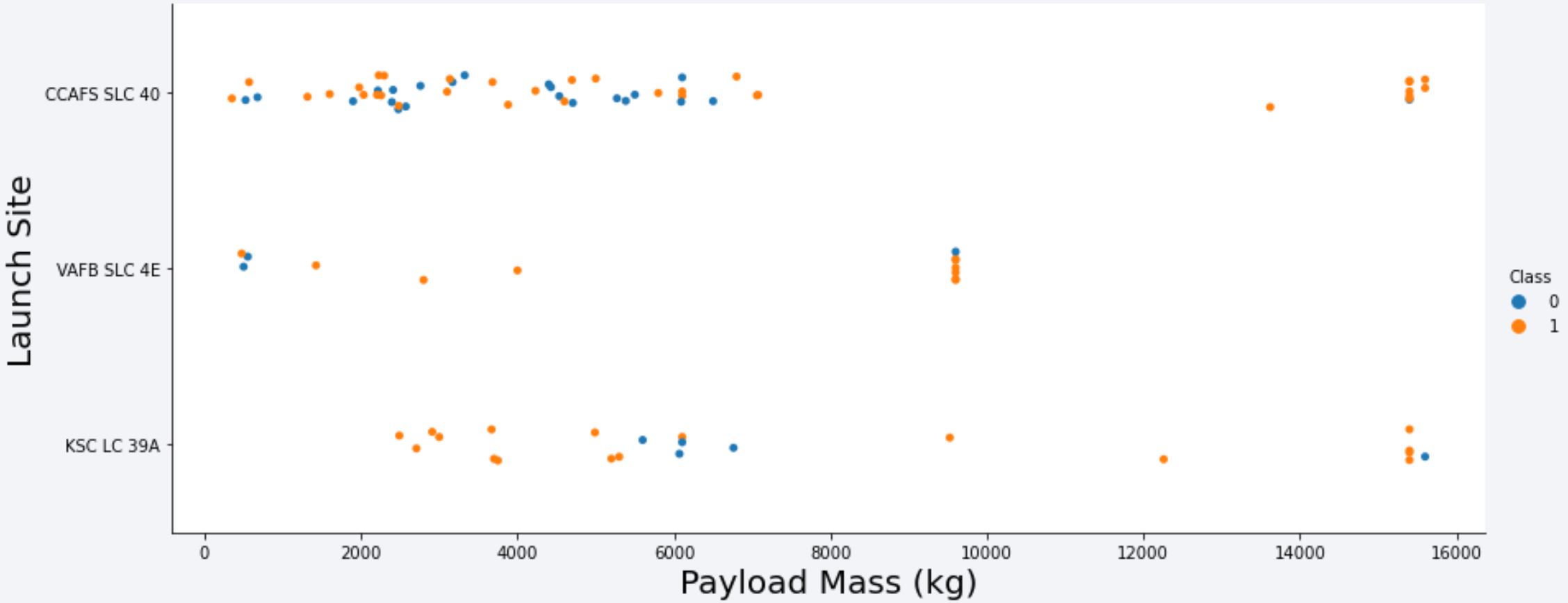

Section 2
Insights drawn
from EDA

Flight Number vs. Launch Site



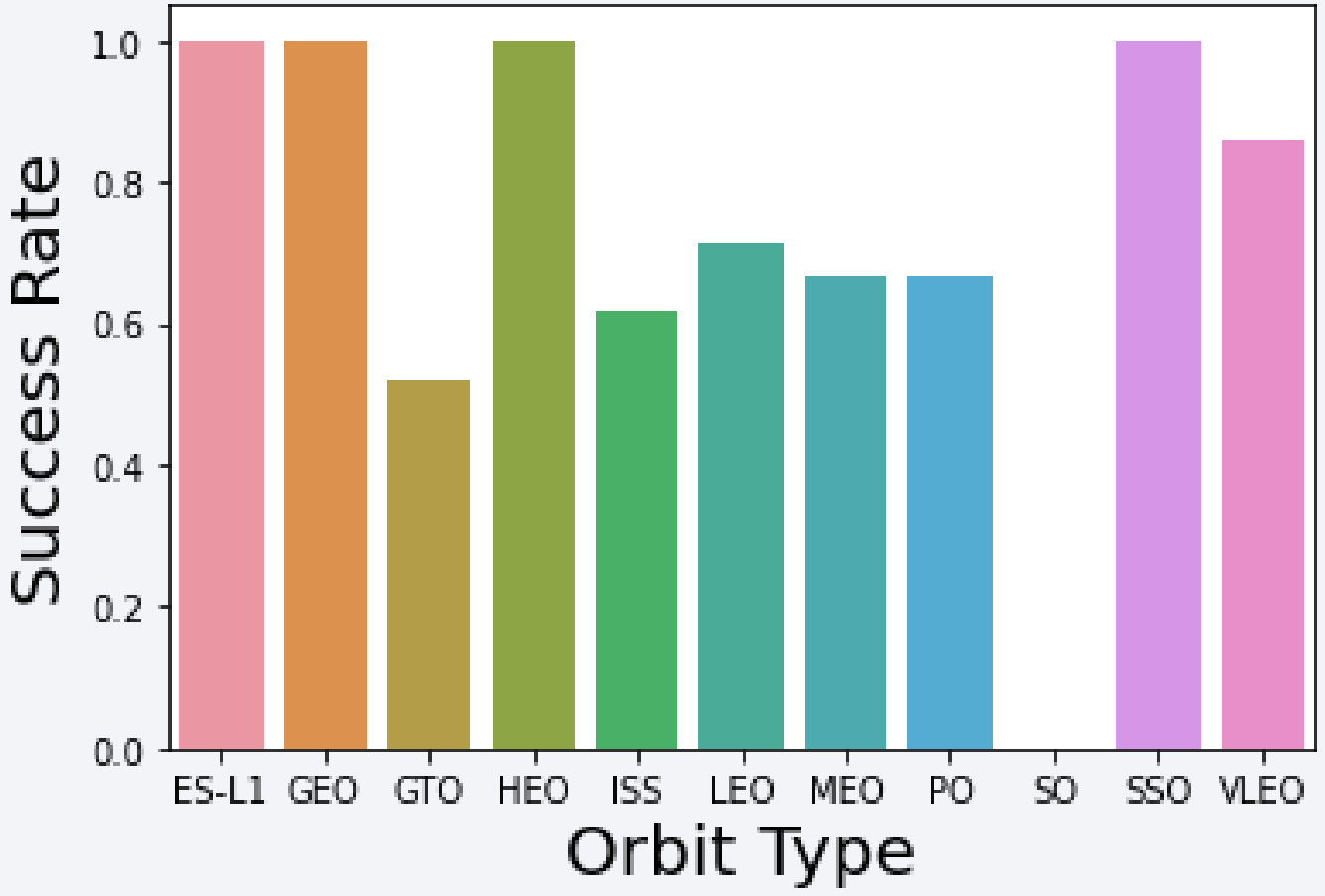
It seams to have a higher success rate at VAFB SLC 4E and KSC LC 39A launch sites compare to CCAFS SLC 40 site.

Payload vs. Launch Site



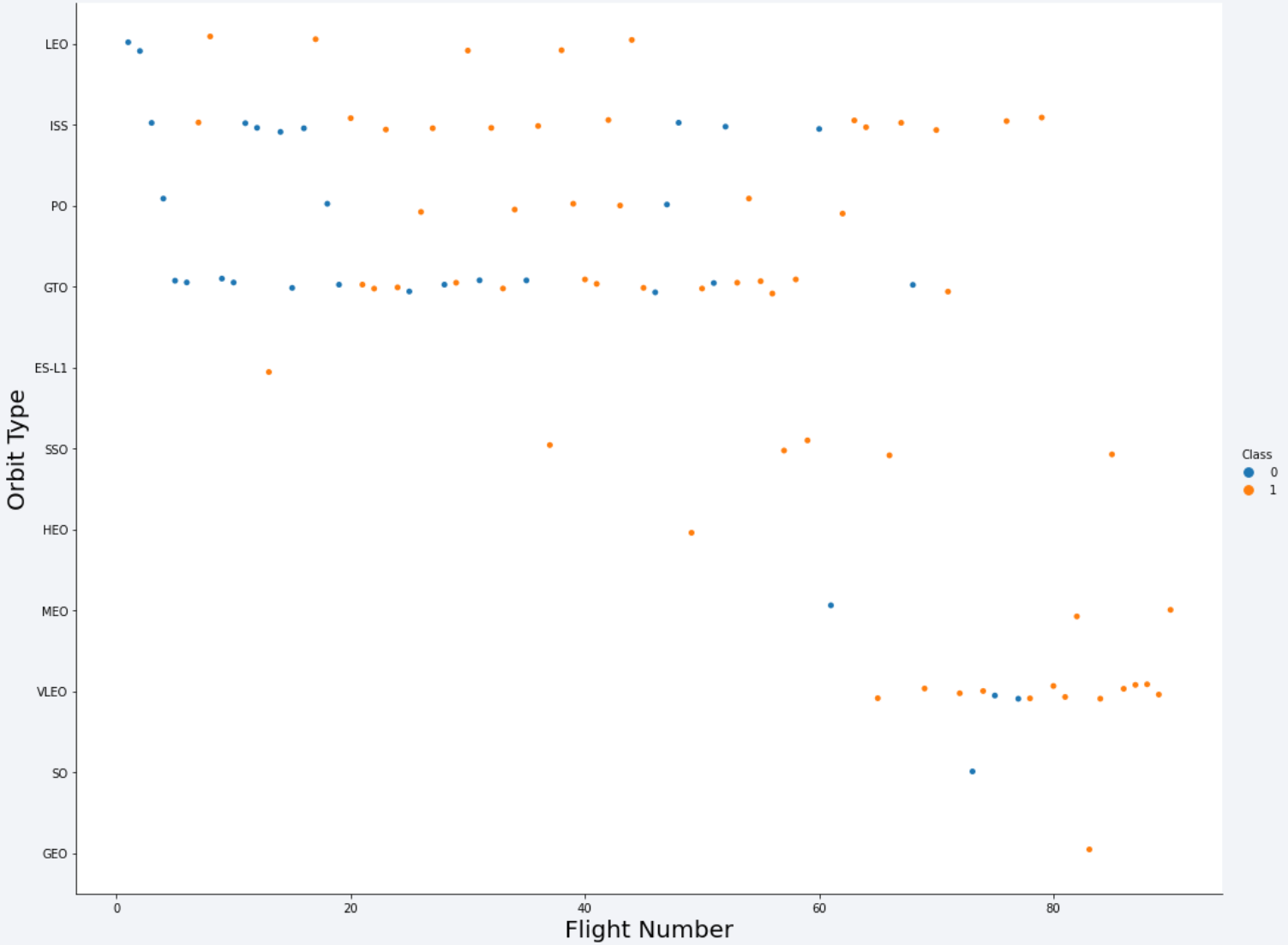
At the VAFB-SLC launch site there are no rokets launches for heavy payload mass, greater than 10000.

Success Rate vs. Orbit Type



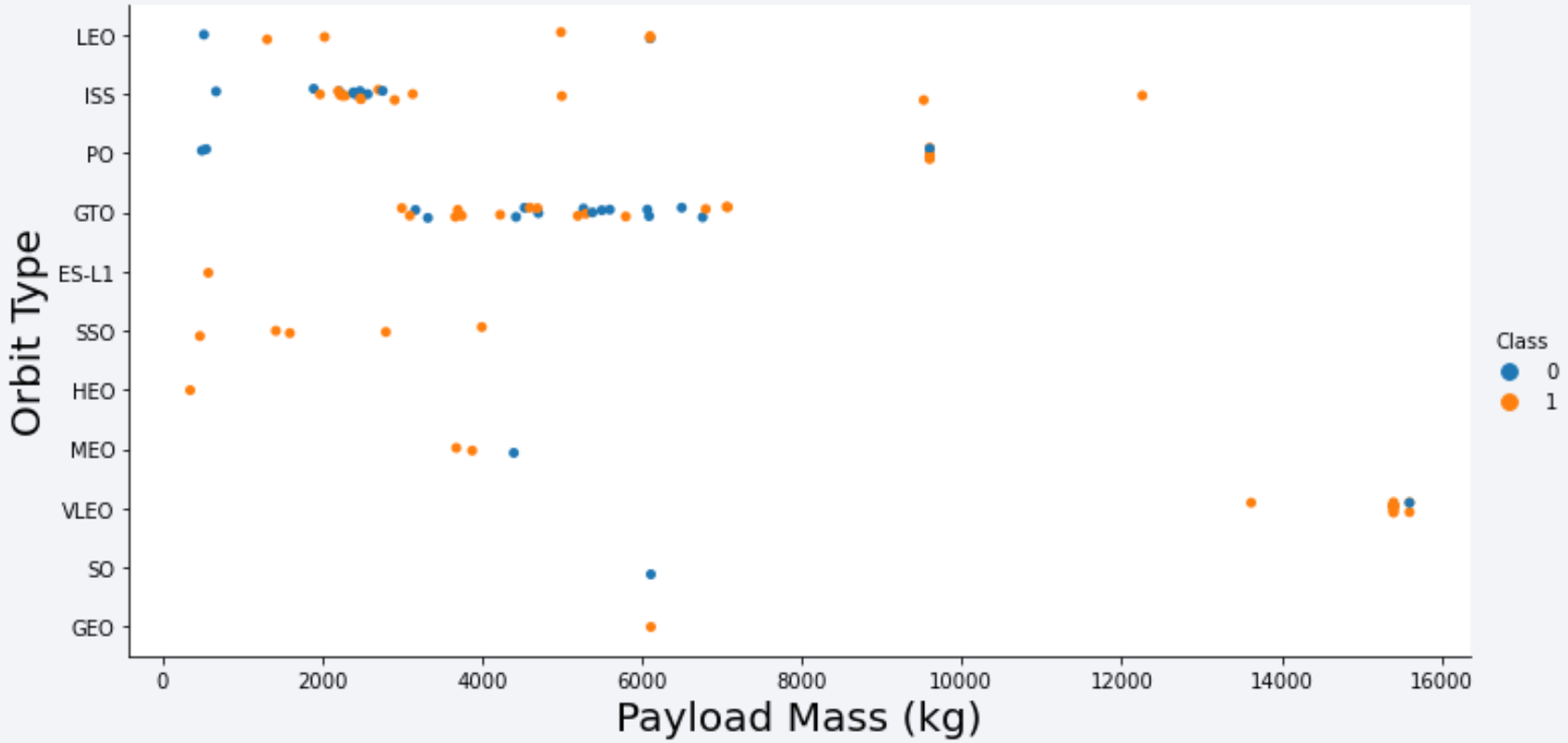
The success rate is higher at the following orbit types:
ES-L1, GEO, HEO, and SSO.

Flight Number vs. Orbit Type



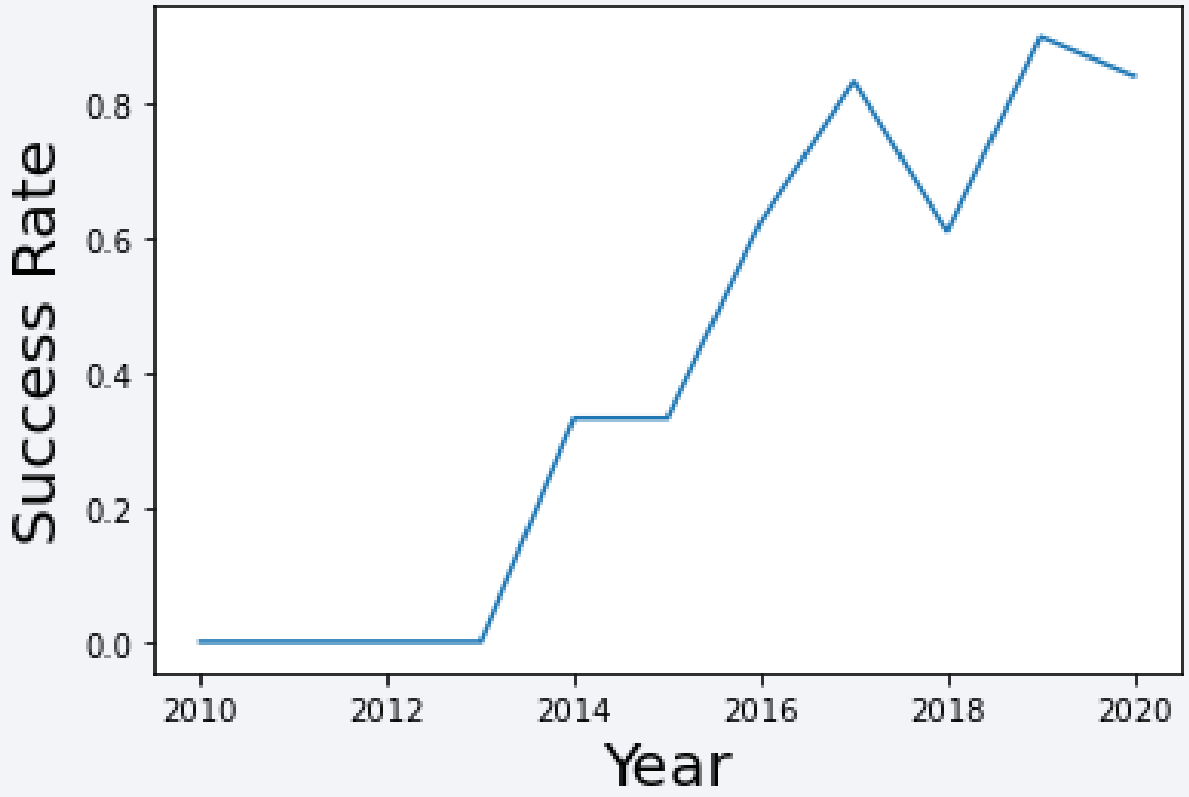
In the LEO orbit the success is related to the flight number, however for the GTO orbit it seems to be no relationship with a certain flight number.

Payload vs. Orbit Type



For heavy payloads (excepting for GTO) the successful landing rate are more for SSO, LEO, and ISS.

Launch Success Yearly Trend



The success rate since 2013 kept increasing till 2020.

All Launch Site Names

Find the names of the unique launch sites.

```
%sql SELECT DISTINCT Launch_Site FROM SPACEXTBL;
 * sqlite:///my_data1.db
Done.
```

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40
None

Launch Site Names Begin with 'CCA'

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

```
%sql SELECT * from SPACEXTBL WHERE Launch_Site LIKE 'CCA%' LIMIT 5;
```

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

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
06/04/2010	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0.0	LEO	SpaceX	Success	Failure (parachute)
12/08/2010	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0.0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
22/05/2012	7:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525.0	LEO (ISS)	NASA (COTS)	Success	No attempt
10/08/2012	0:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500.0	LEO (ISS)	NASA (CRS)	Success	No attempt
03/01/2013	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677.0	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

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

```
%%sql SELECT SUM(PAYLOAD_MASS__KG_) TOTAL_PAYLOAD_MASS, Customer FROM SPACEXTBL GROUP BY Customer
    HAVING Customer == 'NASA (CRS)';
```

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

TOTAL_PAYLOAD_MASS	Customer
45596.0	NASA (CRS)

Average Payload Mass by F9 v1.1

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

```
%%sql SELECT AVG(PAYLOAD_MASS__KG_) FROM (SELECT PAYLOAD_MASS__KG_ FROM SPACEXTBL GROUP BY Booster_Version
                                          HAVING Booster_Version LIKE 'F9 v1.1%')
```

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

AVG(PAYLOAD_MASS__KG_)
2413.4545454545455

First Successful Ground Landing Date

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

```
%sql SELECT date, Landing_Outcome FROM SPACEXTBL WHERE Landing_Outcome = 'Success (ground pad)' LIMIT 1;
```

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

Date	Landing_Outcome
22/12/2015	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.

```
%%sql SELECT DISTINCT Booster_Version FROM SPACEXTBL WHERE Landing_Outcome = 'Success (drone ship)' and
    PAYLOAD_MASS__KG_ > 4000 and PAYLOAD_MASS__KG_ < 6000;
```

```
 * 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

Calculate the total number of successful and failure mission outcomes.

```
%%sql SELECT COUNT( CASE WHEN Mission_Outcome LIKE '%Failure%' THEN 1 END ) Failure,
       COUNT ( CASE WHEN Mission_Outcome LIKE '%Success%' THEN 1 END ) Success FROM SPACEXTBL
```

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

Failure	Success
1	100

Boosters Carried Maximum Payload

```
%%sql SELECT Booster_Version FROM SPACEXTBL WHERE PAYLOAD_MASS__KG_ =
    (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL)
```

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

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

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

2015 Launch Records

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

```
%%sql SELECT substr(Date, 4, 2), Landing_Outcome, Booster_Version, Launch_Site FROM SPACEXTBL
    WHERE Landing_Outcome LIKE 'Failure%' and substr(Date, 7, 4) = '2015'
```

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

substr(Date, 4, 2)	Landing_Outcome	Booster_Version	Launch_Site
10	Failure (drone ship)	F9 v1.1 B1012	CCAFS LC-40
04	Failure (drone ship)	F9 v1.1 B1015	CCAFS LC-40

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.

```
%%sql SELECT Landing_Outcome, COUNT(Landing_Outcome) rank FROM SPACEXTBL
    WHERE (Landing_Outcome = 'Failure (drone ship)' OR Landing_Outcome = 'Success (groung pad)')
    AND (DATE BETWEEN '04/06/2010' and '20/03/2017')
    ORDER BY rank DESC
```

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

Landing_Outcome	rank
Failure (drone ship)	3

Section 3

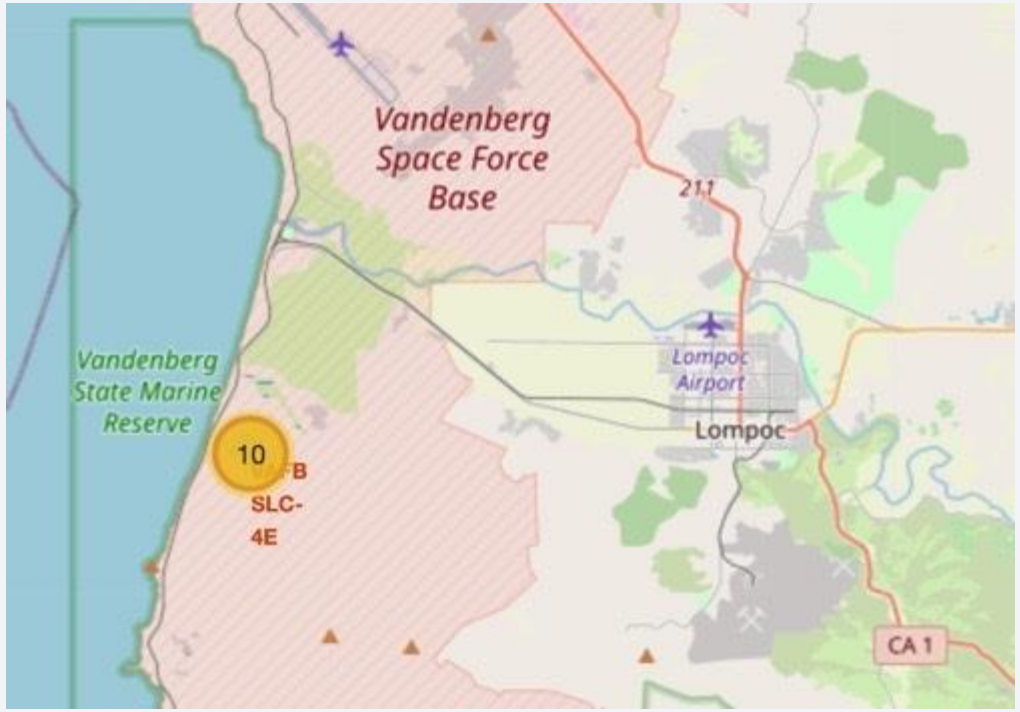
Launch Sites Proximities Analysis

Launch Sites' Locations

There are four Launch Site locations – two on the West Coast and two on the East Coast of US: CCAFS LC-40, CCAFS SLC-40, KSC LC-39A (in Florida), and VAFB SLC-4E (in California).

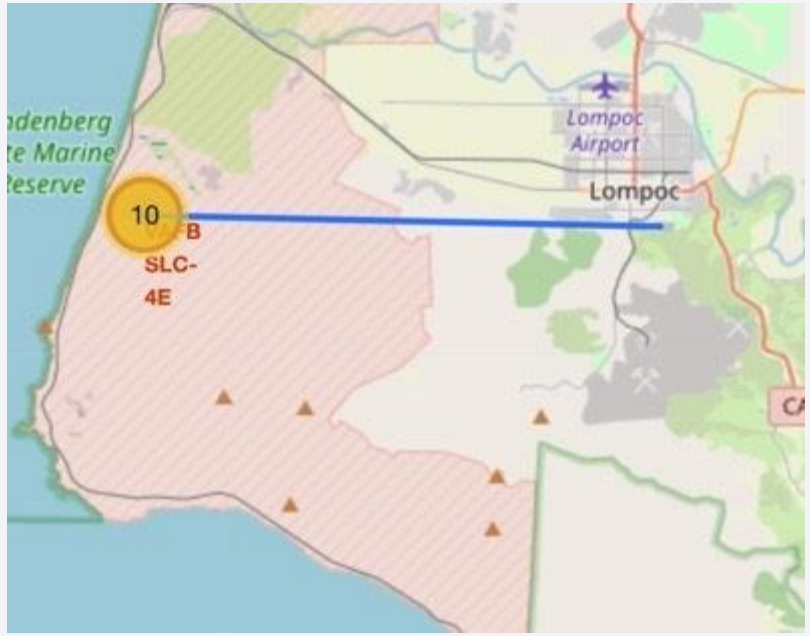
Proximities to the Launch Sites

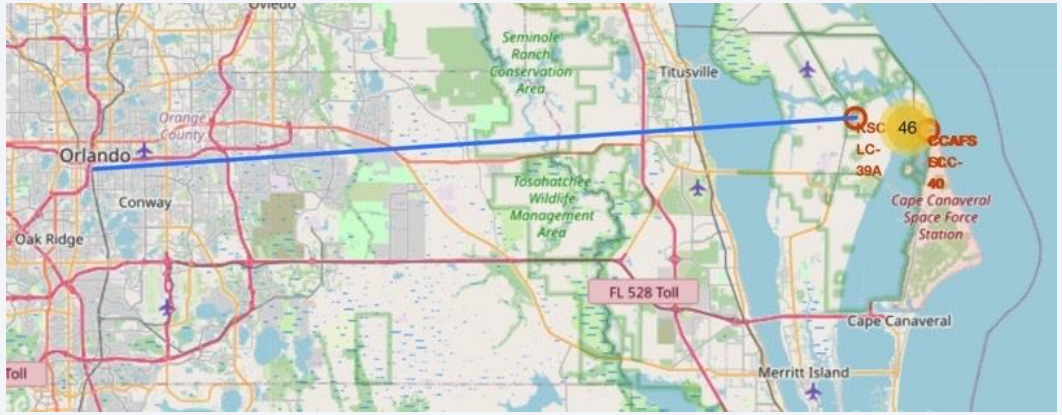
All launch sites are close to the coast. The major cities closer to a site are Orlando on the East Coast, and Lampoc on the West Coast.



Distance to a major location

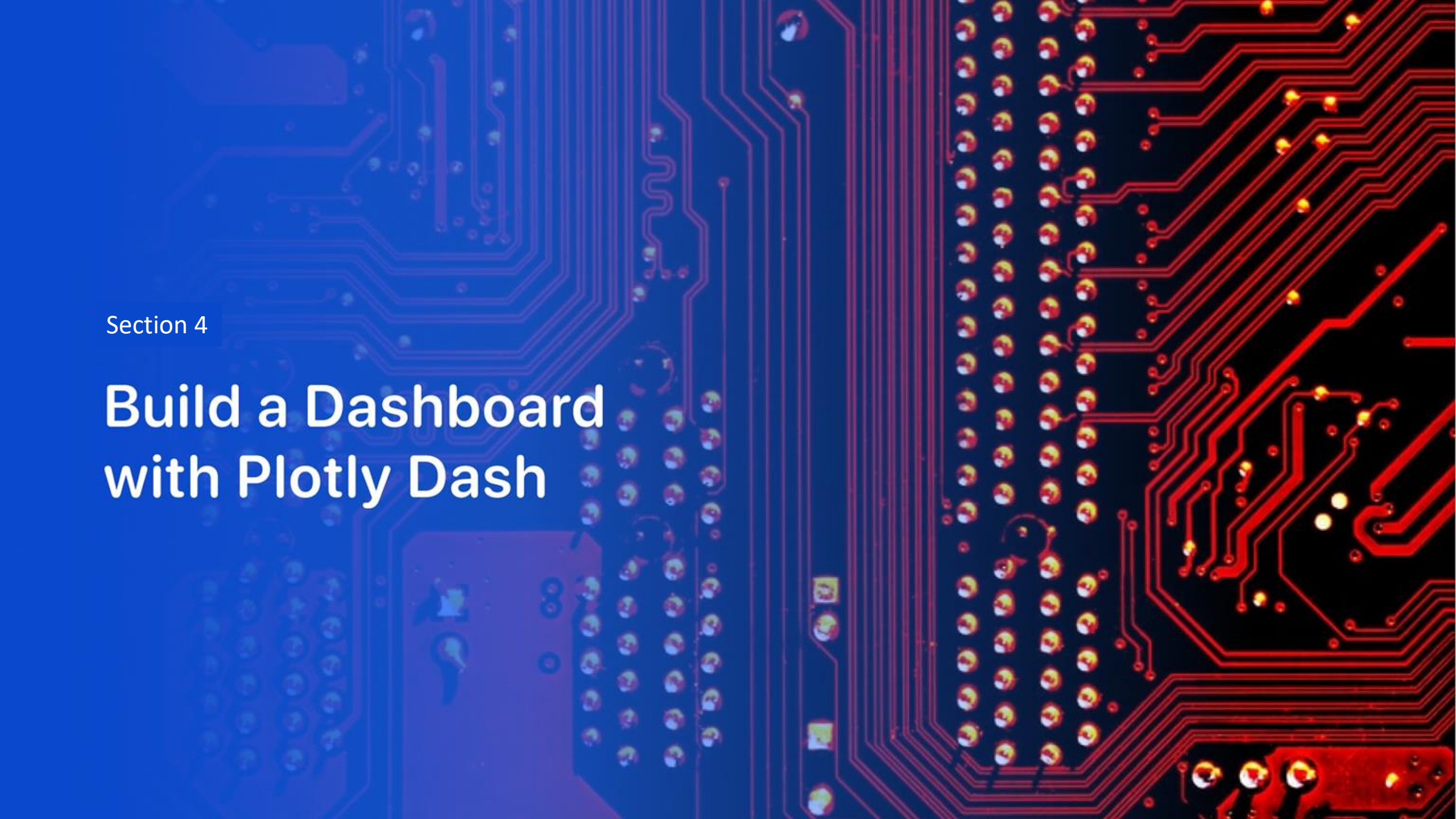
All of the locations are very close to the coast and each of them are in a isolated unpopulated place, for successful landing.





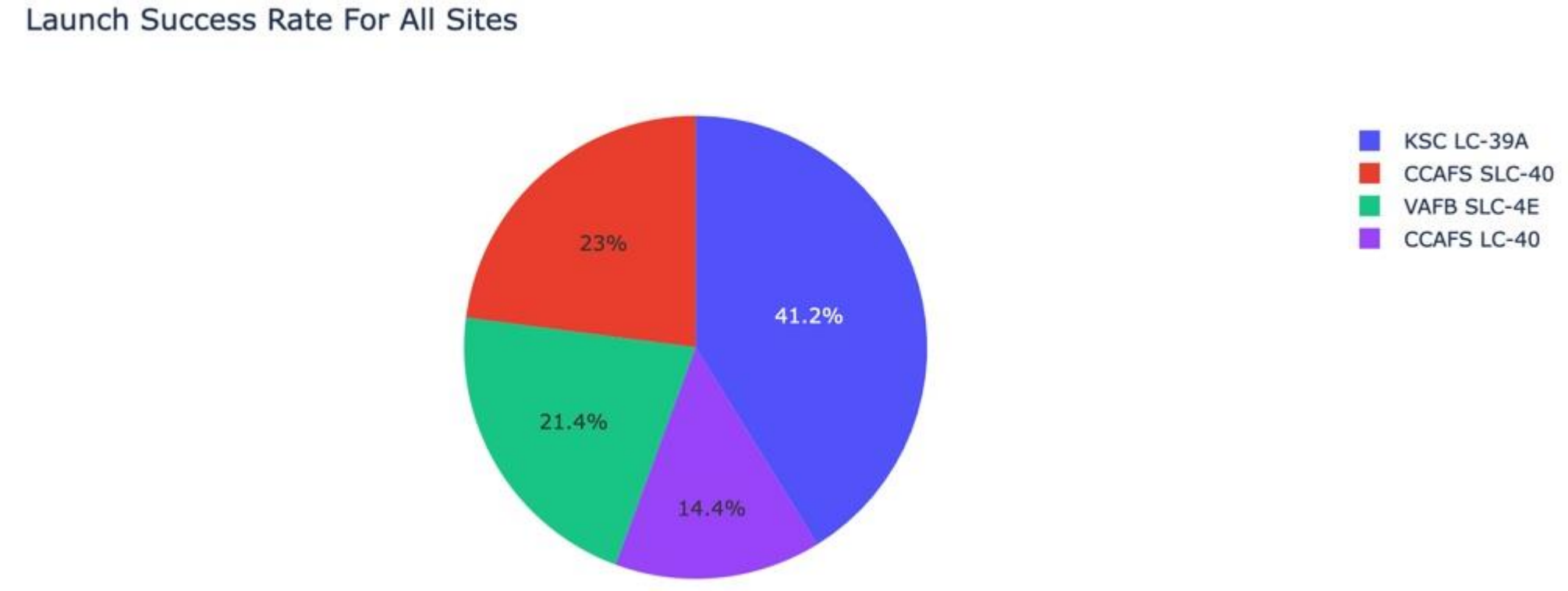
Section 4

Build a Dashboard with Plotly Dash



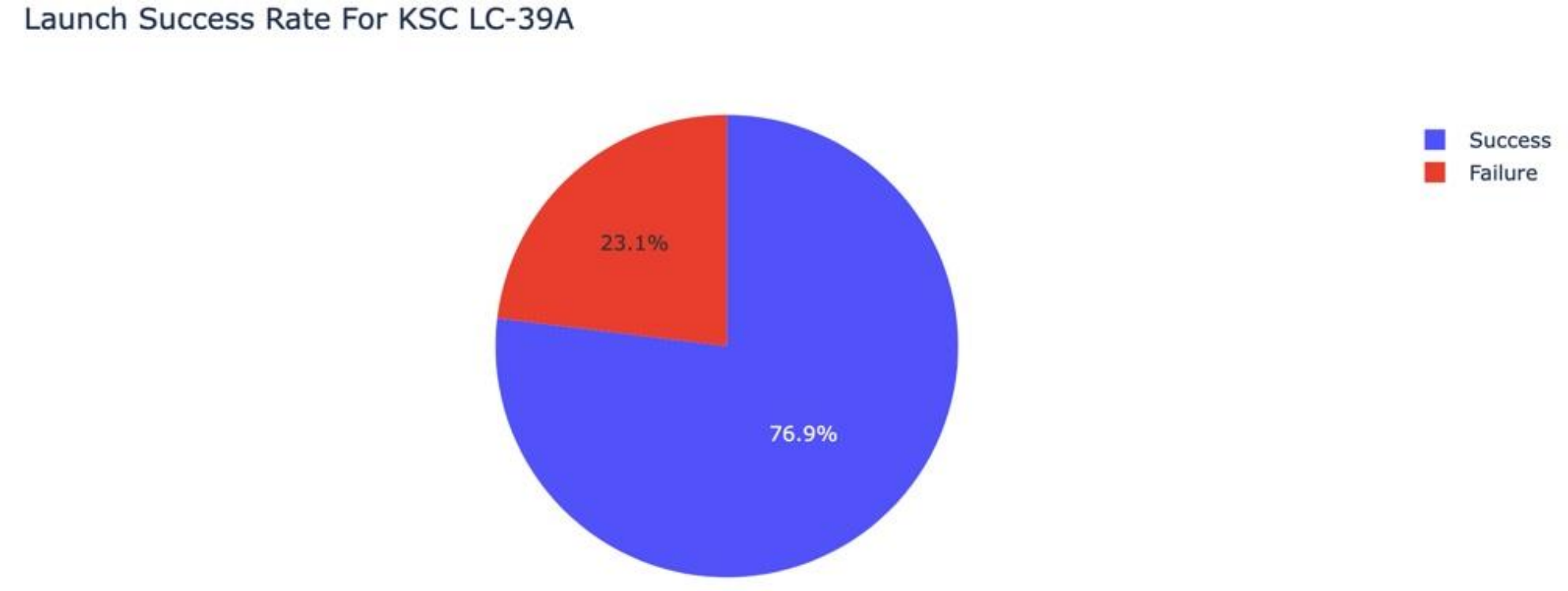
Launch Success count for all sites

The sites situated on East Cost are on the lead: KSC LC-39A with 41.2%, followed by CCAFS SLC-40 site with 23%.



The site with highest success ratio

KSC LC-39A has the highest success ration of 76.9%.



Launch Outcome with different payload

0 – 5000 Payload Mass

Launch Outcome with different payload

2000 – 6000 Payload Mass

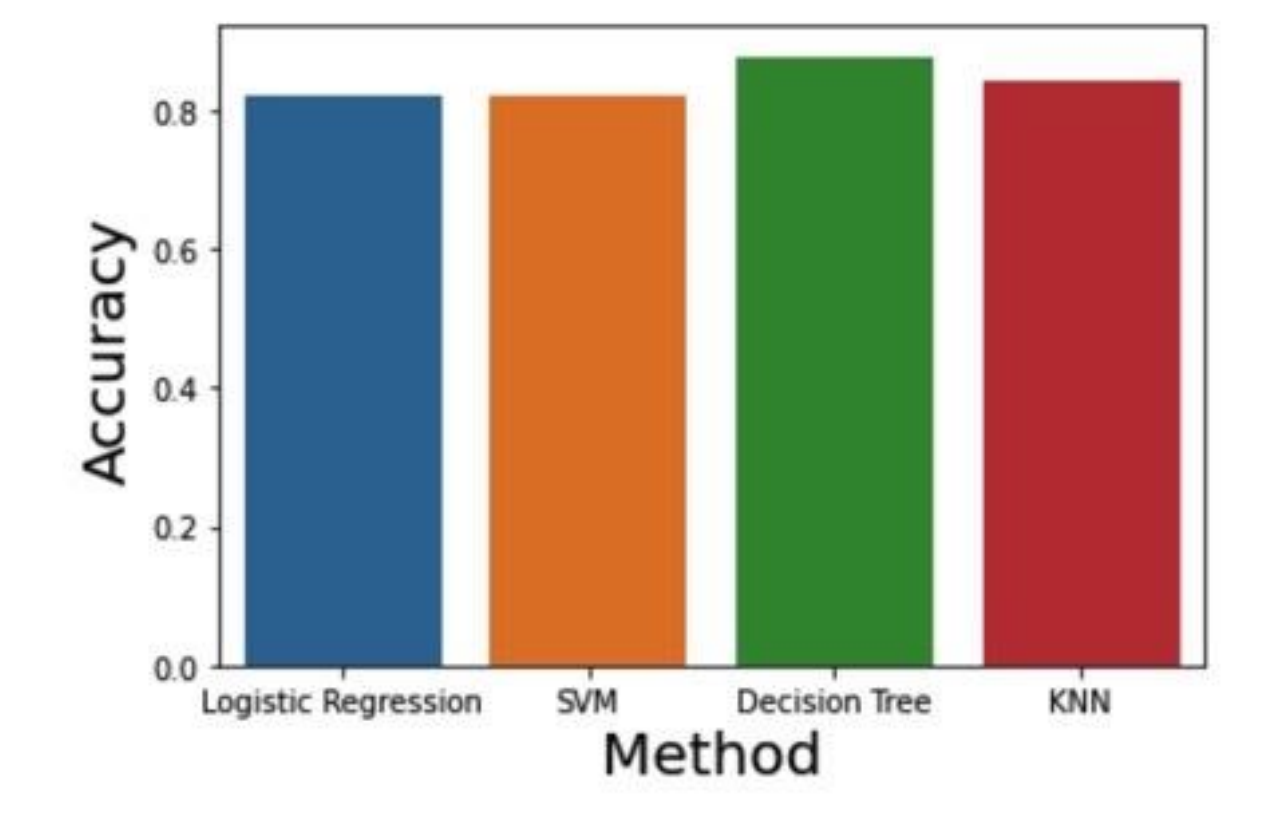
Launch Outcome with different payload

5000 – 10000 Payload Mass

Section 5

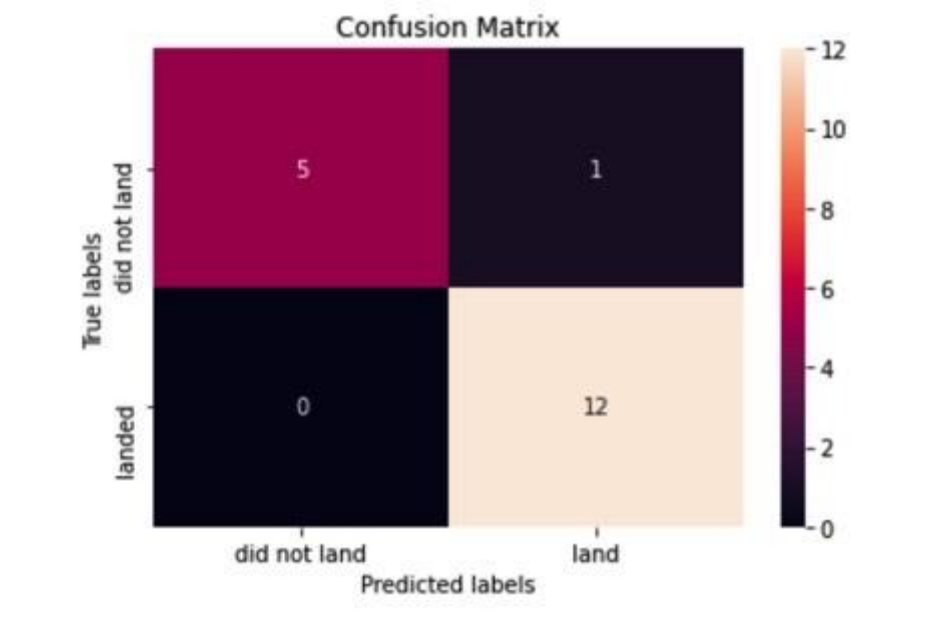
Predictive Analysis (Classification)

Classification Accuracy



The model with the best accuracy is Decision Tree.

Confusion Matrix



The confusion matrix of Decision Tree model shows 12 successful landings and 5 that did not land, with only one outcome as being misclassified.

Conclusions

- CCAFS LC-40, VAFB SLC-4E, KSC LC-39A, and CCAFS SLC-40 are the launch sites in the space mission.
- Using Decision Tree we can accurately predict successful landings at each site.
- For an successful landing each site should be situated very close to the coast (preferably on east), isolated from major locations.

Appendix

- The Python code notebooks, SQL queries, charts created during this project can be found on https://github.com/mbujac/Applied-Data-Science-Capstone/blob/ca579780d9b7a7e2ffd9f0f616e7bda7cf9c228a/README.md

Thank you!